PROVISIONAMENTO DE RECURSOS E QOS EM REDES DE NÚCLEO IP PARA SISTEMAS CELULARES DE 3^a GERAÇÃO

Saulo Vaz de Vasconcellos

TESE SUBMETIDA AO CORPO DOCENTE DA COORDENAÇÃO DOS PROGRAMAS DE PÓS-GRADUAÇÃO DE ENGENHARIA DA UNIVERSIDADE FEDERAL DO RIO DE JANEIRO COMO PARTE DOS REQUISITOS NECESSÁRIOS PARA A OBTENÇÃO DO GRAU DE MESTRE EM CIÊNCIAS EM ENGENHARIA ELÉTRICA.

Aprovada por:

Prof. José Ferreira de Rezende, Dr.

Prof. Aloysio de Castro Pinto Pedroza, Dr.

Prof. Djamel Fawazi Hadj Sadok, Ph.D.

RIO DE JANEIRO, RJ - BRASIL

SETEMBRO DE 2002

VASCONCELLOS, SAULO VAZ

Provisionamento de Recursos e QoS em redes de Núcleo IP para Sistemas Celulares de 3^a Geração [Rio de Janeiro] 2002

XIV, 104 p. 29,7 cm (COPPE/UFRJ, M.Sc., Engenharia Elétrica, 2002)

Tese - Universidade Federal do Rio de Janeiro, COPPE

1. Redes Móveis
2. Qualidade de Serviço
3. Avaliação de Desempenho

I. COPPE/UFRJ II. Título (série)

Aos meus tios, Ceiça e Weimar.

Agradecimentos

À Deus, por nos dar vida e consciência.

Aos meus pais Luiz e Joselina, pelo amor, carinho, apoio, enfim, por tudo.

A todos os meus familiares, que sempre se interessaram e me apoiaram, me ajudando nessa caminhada.

Ao Prof. José Rezende, pela amizade, apoio e orientação do trabalho.

Aos Profs. Aloysio e Djamel, pela presença na banca e pelas sugestões ao texto final.

Aos demais professores do GTA, Leão e Otto.

À Poli, por estar ao meu lado, apoiando-me com carinho e paciência.

Ao amigo e colega Kleber, que ajudou o trabalho de diversas formas, nas pesquisas bibliográficas iniciais e dando opiniões sobre seu andamento.

Aos amigos e colegas da COPPE. São muitos, mas vou tentar me lembrar deles: Valentim, Roman, Gardel, Fagundes, Rubi, Eric, Pedro, Doc, Bernardo, Márcio, Marcial, Granato, Belem, Mauros, David, Bagatelli, Aline, Paulo, Artur, Luís, Baiano, André, Roberta, Roberta Guinther, Ivana, LG, Sidney, Glauco, do GTA; ao pessoal da sistemas: Isaac, Beto e Guto; ao Augusto do LPS. Aos amigos da Alegrete e à “Galera do Bronx”, pela paciência e pela força.

À Bia, Solange e Wilson pelo suporte operacional.

Ao PEE/COPPE/UFRJ, pelas instalações e equipamentos utilizados.

À CAPES, pelo financiamento da pesquisa.

Resumo da Tese apresentada à COPPE/UFRJ como parte dos requisitos necessários para a obtenção do grau de Mestre em Ciências (M.Sc.)

PROVISIONAMENTO DE RECURSOS E QOS EM REDES DE NÚCLEO IP PARA SISTEMAS CELULARES DE 3^a GERAÇÃO

Saulo Vaz de Vasconcellos

Setembro/2002

Orientador: José Ferreira de Rezende

Programa: Engenharia Elétrica

Esse trabalho apresenta um estudo de qualidade de serviço (QoS) em Redes de Núcleo IP de sistemas celulares de terceira geração. Inicialmente é realizado um levantamento de características das redes móveis que influenciam a QoS. São então apresentadas as principais soluções de QoS para redes IP fixas. Elas são analizadas e comparadas de forma crítica. Suas deficiências para lidar com ambientes de mobilidade são discutidas.

A partir da observação de características das aplicações móveis e das necessidades de seus usuários, são propostas classes de serviço voltadas para mobilidade. Essas classes procuram flexibilizar o suporte à determinados parâmetros de QoS para garantir requisitos fundamentais. Propõe-se a utilização de três diferentes mecanismos para a implementação de uma das classes sugeridas, a de tempo real com mobilidade. Dois destes são escalonadores tradicionalmente empregados para diferenciação de serviços (WRR e PQ) e o terceiro é um mecanismo de gerenciamento de *buffer* baseado na política protetora SPP. É avaliada a interação desses mecanismos com propostas de controle de tráfego (policiamento e MBAC). Resultados de simulações mostram que o mecanismo de SPP é o mais adequado para dar suporte a classe em questão. Ele foi capaz de oferecer as garantias necessárias, respeitando as especificações da classe.

Abstract of Thesis presented to COPPE/UFRJ as a partial fulfillment of the requirement for the degree of Master of Science (M.Sc.)

RESOURCE PROVISIONING AND QOS IN IP CORE NETWORKS FOR 3G CELLULAR SYSTEMS

Saulo Vaz de Vasconcellos

September/2002

Advisor: José Ferreira de Rezende

Department: Electrical Engineering

This work presents a study of Quality of Service (QoS) in IP Core Networks for 3G cellular systems. A study of mobile environments and its characteristics related to QoS is carried through. The traditional QoS solutions for fixed IP networks are presented. These solutions are critically analyzed and compared to each other. Its deficiencies in dealing with mobility are pointed out.

New service classes are proposed, based on mobile applications characteristics and on users requirements. These classes offer flexibility in secondary QoS parameters in order to guarantee the most important ones. It is suggested the use of three mechanisms for supporting one of the proposed classes: the real time multimedia with mobility. Two of these mechanisms are schedulers, often used in differentiated services environments (PQ and WRR), and the third is a buffer management based on a protective policy, the SPP. The inter-operation of these mechanisms with traffic control (policing and MBAC) is evaluated. Simulation results show that the SPP mechanism is better than the others to implement the proposed class. It was capable of offering the necessary QoS guarantees described in the class specification.

Lista de Acrônimos

2G : *Second Generation*;

3G : *Third Generation*;

3GPP : *Third Generation Partnership Project*;

AAA : *Authentication, Authorization and Acounting*;

AF : *Assured Forwarding*;

AMPS : *Advanced Mobile Phone System*;

ATM : *Asynchronous Transfer Mode*;

BS : *Base Station*;

CDMA : *Code Division Multiple Access*;

CN : *Core Network*;

CODEC : *Coder-Decoder*;

DiffServ : *Differentiated Services*;

DSCP : *DiffServ Codepoint*;

EF : *Expedited Forwarding*;

FIFO : *First In, First Out*;

FTP : *File Transfer Protocol*;

GGSN : *Gateway GPRS Support Node*;

GPRS : *General Packet Radio Service*;

GSM : *Global System for Mobile Communication*;

IETF : *Internet Engineering Task Force*;

IntServ : *Integrated Services*;

IP : *Internet Protocol*;

MBAC : *Measurement Based Admisson Control*;

MOS : *Mean Opinion Score*;

MS : *Mobile Station*;

ns-2 : *Network Simulator 2*;

OTcl : *Object Tool command language*;

PBAC : *Parameter Based Admisson Control*;

PCM : *Pulse-Code Modulation*;

PDB : *Per Domain Behavior*;

PDR : *Per Domain Reservation*;

PHB : *Per Hop Behavior*;

PHR : *Per Hop Reservation*;

PQ : *Priority Queue*;

QoS : *Quality of Service*;

RED : *Random Early Detection*;

RMD : *Resource Management in DiffServ*;

RNC : *Radio Network Controller*;

RNS : *Radio Network System*;

RODA : *Resource Management in DiffServ on Demand*;

RSVP : *Resource Reservation Protocol*;

SGSN : *Serving GPRS Support Node*;

SLA : *Service Level Agreement*;

SPP : *Simulated Protective Policy*;

TCP : *Transmission Control Protocol*;

TDMA : *Time Division Multiple Access*;

UDP : *User Datagram Protocol*;

UMTS : *Universal Mobile Telecommunications System*;

UTRAN : *UMTS Terrestrial Radio Access Network*;

VINT : *Virtual Internetwork Testbed*;

WRR : *Weighted Round-Robin.*

Sumário

Lista de Figuras

Lista de Tabelas

Capítulo 1

Introdução

O MUNDO das telecomunicações vem sofrendo grandes mudanças nos últimos anos. A disponibilização das informações ganhou extrema facilidade a partir do extraordinário crescimento da Internet, a rede mundial de computadores. Para que se possa obter uma completa integração da transmissão de diferentes mídias através da Internet, como voz e vídeo, juntamente com dados, é necessário que se realizem modificações em sua proposta inicial. Tais modificações incluem o suporte a diferentes níveis de Qualidade de Serviço (QoS - *Quality of Service*), o que irá permitir que as diferentes aplicações sejam respeitadas em suas características fundamentais.

Outro setor das telecomunicações que tem popularidade crescente é a área da comunicação móvel. A telefonia celular foi o primeiro passo em direção à comunicação pessoal móvel. A Internet e as comunicações móveis são os negócios em telecomunicações que vêm apresentando o maior crescimento nos últimos anos [1]. A disponibilização no mercado de PDA's (*Personal Digital Assistants*) e celulares "inteligentes" a preços cada vez mais acessíveis, traz serviços atraentes pela extrema flexibilidade. Assim, é natural que ocorra uma integração entre as duas áreas. Usuários móveis inicialmente só podiam utilizar as redes celulares para a transmissão de tráfego de voz através de serviços de telefonia. Atualmente eles já têm acesso a alguns serviços de Internet através das infra-estruturas atualmente em operação, como envio de pequenas mensagens por *e-mail*, ou consulta de *sites* simplificados especialmente construídos. A próxima geração de redes celulares, a dita terceira geração (*third generation* - 3G), irá contar com serviços mais complexos, como a transmissão de vídeo através da rede. Os custos dos serviços vêm baixando e usuários já chegam, em alguns casos, a optar pela comodidade da conectividade oferecida nos serviços móveis. Com o surgimento de redes de pacotes sem fio, os custos tendem a diminuir ainda mais, uma vez que os recursos disponíveis passam a ser melhor aproveitados.

As redes celulares de terceira geração devem então oferecer serviços com QoS assegurada fim-a-fim, para dar o suporte necessário às novas aplicações. Para tanto, todos os componentes dos sistemas celulares devem ser capazes de garantir níveis mínimos de QoS para que o serviço possa ser composto. Um dos elementos fundamentais das arquiteturas 3G é a Rede de Núcleo. Essas Redes de Núcleo são responsáveis pelo transporte de todo o tráfego trocado entre os terminais móveis e as redes fixas, como por exemplo a Internet. Ela realiza também a interconexão entre diversas redes de acesso de rádio da mesma ou de diferentes operadoras que prestam o serviço 3G.

Há uma forte tendência para que essas Redes de Núcleo, assim como toda a tecnologia de transmissão nas redes 3G, sejam construídas a partir de tecnologia IP, o que permitiria uma inter-operação transparente com a Internet. Como colocado anteriormente, as redes IP não foram inicialmente projetadas para oferecer garantias de QoS aos fluxos de dados. Atualmente existem algumas propostas para permitir a oferta de QoS nessas redes. Entretanto, além de lidar com as tradicionais dificuldades em provisão de QoS das redes IP, as Redes de Núcleo devem ainda solucionar algumas questões relacionadas com a mobilidade, como a imprevisibilidade do movimento dos nós. Essa imprevisibilidade pode levar a grandes variações nos níveis de utilização da rede, trazendo maior dificuldade a um provisionamento adequado de QoS que simultaneamente permita altos níveis de utilização dos recursos.

Esse trabalho faz então um estudo de qualidade de serviço para redes móveis, focalizando a qualidade de serviços oferecida pela infra-estrutura fixa das Redes de Núcleo de sistemas celulares. Essas infra-estruturas fixas precisam lidar com algumas características particulares, como altas variações no nível de utilização, o que dificulta uma utilização otimizada da rede.

1.1 Objetivos

Esse trabalho tem como objetivo o estudo das características de sistemas móveis e as possibilidades de construção de um serviço com qualidade fim-a-fim para o usuário final. As soluções propostas devem ser capazes de oferecer serviços com os níveis de QoS necessários às aplicações, sem deixar de possibilitar o atendimento a uma larga escala de usuários e procurando manter, ainda, uma alta utilização dos recursos disponíveis. O trabalho é voltado para as porções fixas dos sistemas móveis, as Redes de Núcleo. Essas redes devem lidar com características de tráfego diferentes daquelas normalmente encontradas nos sistemas fixos.

Dessa forma, é realizada inicialmente uma investigação das principais soluções de QoS para a Internet, observando-se sempre suas capacidades de lidar com os requisitos de QoS das Redes de Núcleo de sistemas 3G. São abordadas as duas principais propostas de QoS para a Internet, a Integração de Serviços (IntServ) [2] e a Diferenciação de Serviços (DiffServ) [3, 4], estudadas no âmbito do *"Internet Engineering Task Force"* (IETF) [5], o principal fórum de discussão de tecnologias para a Internet.

Conclui-se que essas soluções representam paradigmas opostos em termos de provisionamento de recursos. Enquanto o IntServ pode ser visto como uma arquitetura de QoS dinâmica, o DiffServ apresenta-se como uma solução baseada em provisionamento estático de recursos. Cada uma dessas soluções possuem qualidades e defeitos complementares quando se pensa em QoS para ambientes móveis. Apesar de permitir o atendimento a grandes escalas de usuários, o DiffServ pode ter dificuldades em garantir níveis de QoS em ambientes altamente dinâmicos. O IntServ, por outro lado, oferece garantias rígidas de QoS aos fluxos de dados, necessitando entretanto de uma alta carga de sinalização, o que dificulta sua aplicação em redes que lidam com grandes quantidades de usuários como o caso das Redes de Núcleo de sistemas celulares. Algumas propostas híbridas de QoS mais recentes são também discutidas, sempre observando-se as características relevantes à mobilidade. Estudam-se, então, diferentes mecanismos de gerenciamento de recursos utilizados nas arquiteturas descritas, focalizando principalmente o controle da carga, que pode ser realizado de diferentes formas: desde o controle de admissão por chamada até o policiamento de agregações de tráfego, que pode ser visto como um controle de admissão em nível de pacotes.

Face a dificuldade em se oferecer QoS para as Redes de Núcleo dos sistemas celulares através das propostas tradicionais para a Internet, propõe-se a criação de classes de serviços voltadas para a mobilidade. Os ambientes de mobilidade trazem novos parâmetros de QoS que devem ser considerados ao se propor novas classes. Além dos parâmetros já existentes em redes fixas é importante levar em conta a probabilidade de conexão contínua e o perfil de perdas [6, 7]. Além disso, essas novas classes de serviço devem tirar proveito da robustez de aplicações voltadas para ambientes móveis. Assim, essas classes são projetadas de forma a permitir a degradação de parâmetros de QoS de menor importância para poder fornecer maiores garantias aos requisitos fundamentais. Cada aplicação possui níveis diferenciados de importância para cada parâmetro de QoS. Serviços de voz interativos, por exemplo, necessitam de baixos retardos sendo robustos a certos níveis de perdas, devendo possuir baixa possibilidade de desconexão. As classes flexíveis permitem uma melhoria na utilização global da rede, o que se reflete em um maior número de usuários sendo atendidos, e ao mesmo tempo evitam complexos processos de sinalização.

A partir da proposta das classes de serviço, são analisadas possíveis implementações para as mesmas. O estudo é centrado em uma classe serviços de tempo real em ambientes móveis. As aplicações dessa classe priorizam baixos atrasos e baixas probabilidades de desconexão, sendo robustas a certos níveis de degradação de taxa de perdas. Assim, propõe-se a utilização de um mecanismo de gerenciamento de *buffer* nesse contexto, a "Política Protetora Simulada" (*Simulated Protective Policy* - SPP) [8], para a implementação dos serviços dessa classe. Compara-se, a partir de resultados de simulações, a implementação da classe baseada no SPP com implementações baseadas na utilização de escalonadores amplamente empregados em ambientes de diferenciação de serviços, o *Round Robin* com peso (*Weighted Round*

Robin - WRR) [9] e o Fila Prioritária (*Priority Queue* - PQ).

O estudo dos mecanismos de gerenciamento de recursos é então complementado através da avaliação da utilização de um esquema de controle de admissão em conjunto com uma das classes propostas. Assim, é realizado um estudo sobre a comportamento de um controle de admissão baseado em medidas (*Measurement Based Admisson Control* - MBAC) em diferentes níveis de mobilidade. A utilização conjunta do MBAC com os mecanismos de suporte a classe também é avaliada, através de simulações. O MBAC deve permitir um controle eficiente dos níveis de carga impostos à rede.

1.2 Estrutura do texto

O texto aqui apresentado inicia-se com a introdução das principais características das redes móveis relacionadas com a QoS na infra-estrutura fixa desses sistemas. Em seguida, as soluções tradicionais de QoS são avaliadas de forma crítica em suas capacidades de lidar com mobilidade. As propostas de classes de serviços e de utilização de determinados mecanismos para o seu suporte são apresentados na seqüência. Assim, o capítulo 2 apresenta uma visão geral sobre as redes móveis, incluindo as características particulares do ambiente de mobilidade. Os problemas de QoS que surgem em ambientes com mobilidade e que influenciam diretamente o serviço na porção fixa desses sistemas são o principal foco deste capítulo. Uma breve classificação de sistemas móveis é apresentada. Alguns conceitos de sistemas celulares são também introduzidos nesse capítulo.

O capítulo seguinte, descreve as principais soluções de QoS para redes IP que poderiam ser aplicadas às Redes de Núcleo de sistemas 3G. É realizada uma discussão sobre as características dessas soluções e sobre suas vantagens e desvantagens, no que tange aos serviços que devem ser oferecidos pelos ambientes de mobilidade 3G. Ao final desse capítulo são abordados as principais necessidades e objetivos de um gerenciamento de recursos voltado para o provisionamento de QoS em ambientes de mobilidade

O capítulo 4 traz propostas para o provisionamento de QoS nas Redes de Núcleo de sistemas de terceira geração. Essas propostas procuram meios de assegurar QoS mantendo baixos níveis de sinalização e de armazenamento de estados, buscando a manutenção de propriedade escalar, sem deixar de obter altos níveis de utilização dos recursos da rede e sem deixar de atender aos requisitos dos usuários. Assim, são propostas classes de serviços flexíveis voltadas para ambientes de mobilidade. Alguns mecanismos são propostos para dar suporte a essas classes.

Uma avaliação de desempenho dos mecanismos de suporte à classes de serviços propostas são realizadas no capítulo 5. Para tanto foram utilizadas simulações. O simulador de redes e o cenário simulado são primeiramente abordados. Em seguida é realizada uma comparação crítica entre os resultados obtidos.

Por fim, no capítulo 6, são apresentadas as conclusões obtidas ao longo do trabalho e são apontados algumas das possibilidades de trabalhos futuros. Dessa forma buscou-se apresentar um texto estruturado, que introduzisse as características do problema abordado e que em seguida discutisse e propusesse algumas possíveis soluções, realizando, por fim, uma avaliação de desempenho de algumas dessas propostas.

Capítulo 2

Redes Móveis

ESTE capítulo tem como objetivo apresentar uma visão geral sobre redes móveis, apresentando suas características e suas dificuldades, principalmente ao que se refere à qualidade de serviço de rede. Assim, na seção 2.1 é realizada uma breve classificação de sistemas móveis. Em seguida, na seção 2.2 são descritas características particulares de ambientes de mobilidade. A arquitetura de sistemas celulares, assim como sua evolução é apresentada na seção seguinte. Por fim, na seção 2.4, características de qualidade de serviço para redes móveis são abordadas.

2.1 Sistemas de computação móvel

Os sistemas móveis de computação vêm apresentando grande evolução nos últimos anos. Equipamentos portáteis estão contando com cada vez maior poder de processamento, com baterias que permitem autonomias cada vez maiores e com preços cada dia mais acessíveis. A popularidade dos equipamentos móveis, juntamente com o crescimento da disponibilização de conectividade, irá torná-los uma importante fração do total dos pontos de acesso à Internet. Dessa forma, usuários móveis representarão uma demanda de diferentes tipos de mídias eletrônicas, incluindo além dos tradicionais serviços de telefonia toda uma nova gama de possibilidades. Entre essas possibilidades, incluem-se a transmissão em um mesmo meio de dados e de diferentes aplicações de tempo real, como por exemplo videofonia e videoconferências. Para que isso ocorra é necessário que haja o suporte adequado para o transporte dos dados com a qualidade necessária para cada tipo de aplicação.

A utilização de enlaces sem fio é ponto chave para mobilidade. O simples uso de redes sem fio, entretanto, não é sinônimo de mobilidade. O termo "computação sem fio" apenas indica que os sistemas não estão conectados por cabos à infra-estrutura de rede. Dessa forma a utilização dos enlaces sem fio pode permitir diferentes graus de movimentação aos usuários.

Os chamados sistemas nômades [10], por exemplo, permitem que usuários utilizem seus equipamentos portáteis conectados a uma rede, desde que estejam parados em um ambiente que ofereça a conexão. Cada vez que deseje obter uma conexão, o usuário deve realizá-la de forma explícita. As redes locais sem fio, baseadas na transmissão de dados através de interfaces com dispositivos infra-vermelhos ou de rádio permitem esse tipo de conexão. Na realidade, esse tipo de serviço pode ser oferecido até mesmo em redes convencionais com fio. A mobilidade não é transparente e durante a conexão o usuário utiliza o equipamento de forma estática. Esse seria o caso em que um *laptop* é carregado durante uma viagem, sendo conectado em um ponto de rede (com fio ou não) dentro do quarto do hotel e posteriormente na sala de reuniões no escritório. Apesar de haver mobilidade, o equipamento não pode se comunicar e se mover de forma simultânea.

Já nos sistemas móveis [11], os usuários fazem uso de enlaces sem fio para que possam se mover de forma transparente durante a comunicação de dados. A infra-estrutura utilizada para dar suporte ao usuário móvel, que muitas vezes é bastante complexa, deve ser transparente para a aplicação. Um exemplo popular de um sistema móvel, ainda que com uma variedade reduzida de aplicações, são os sistemas de telefonia celular.

As redes móveis podem ser ainda classificadas em redes com ou sem infra-estrutura. Em redes sem infra-estrutura, os nós móveis se comunicam diretamente entre si, sem haver a necessidade de um ponto centralizado para que se realizem as transmissões (uma infra-estrutura fixa de rede nem sempre é utilizada). Essas redes são chamadas de "redes *Ad hoc*" [12]. As redes com infra-estrutura, por sua vez, se baseiam em um ponto de acesso fixo que coordena toda a comunicação entre os nós móveis, funcionando também como um portal (*gateway*) para redes fixas, como a Internet. Essas redes sem fio funcionam geralmente como redes de acesso à uma infra-estrutura fixa. As redes celulares, apesar de serem redes que contam com infra-estruturas fixas, não devem ser vistas somente como redes de acesso. As redes celulares atualmente existentes possuem tamanhos consideráveis e muitas transmissões podem ocorrer entre pontos dentro da própria rede, sem utilização de infra-estruturas externas. Por outro lado, muitos serviços necessitam acessar pontos externos à rede celular. Assim as redes celulares devem ser consideradas como redes com infra-estrutura fixa, mas que não devem ser classificadas somente como redes de acesso.

2.2 Características

As redes móveis possuem algumas características particulares, muitas vezes sem paralelos em redes fixas. Assim, novos problemas surgem em um ambiente de mobilidade.

Uma característica importante que é observada nas redes móveis é que as faixas de freqüências disponíveis para a transmissão via rádio são bastante estreitas, levando a baixas taxas de transmissão nas redes móveis, quando comparadas às redes fixas tradicionais [13]. É razoável assumir que as taxas para transmissão em enlaces sem fio continuarão cerca de duas ordens de grandeza abaixo das taxas para redes fixas[1], por maior que seja a evolução das tecnologias de transmissão em redes sem fio [11].

Além de mais reduzida, a vazão observada nas redes móveis apresenta uma maior variação do que àquelas observadas em redes fixas. A qualidade e intensidade do sinal podem se modificar com freqüência nos ambientes de mobilidade. Assim a qualidade

[1] As taxas de transmissão em redes locais sem fio chegam à 54 Mbps enquanto que redes locais com fio atingem 1 Gbps.

da transmissão e, conseqüentemente, a taxa máxima observada ficam deterioradas. Outro fator variável que afeta a vazão máxima obtida por cada estação móvel é o número total de usuários transmitindo simultaneamente em uma mesma área. Um grande número de usuários competindo pela transmissão pode degradar a vazão máxima obtida, não só por causa da interferência causada mas também, em alguns casos, por causa da maior disputa pelo meio de transmissão compartilhado. Essas situações de sobrecarga causadas pelo movimento dos terminais podem ocorrer nos pontos de acesso, assim como no interior da infra-estrutura fixa, conforme será discutido na seção 2.2.1.

Como as comunicações são realizadas através de rádio, o ambiente tem grande influência na qualidade da comunicação. As variações na qualidade do canal são causadas pelos mais diversos motivos: ruídos e interferências, problemas de múltiplos percursos, reflexões ou até mesmo problemas climáticos [14, 15]. Em casos extremos, podem ocorrer desconexões, cujas durações podem apresentar grandes variações. Assim, os ambientes de mobilidade apresentam a característica de possuírem taxas de erro binário elevadas, em oposição ao que ocorre em redes fixas [16]. Uma das implicações dessas taxas de erro elevadas é que, ao contrário do caso das redes fixas, as perdas de dados não podem ser consideradas sinais implícitos de congestionamentos, uma vez que as taxas de erro binário não são desprezíveis [17]. Um protocolo que utiliza tal sinalização implícita e que é amplamente aplicado na Internet é o TCP [18], que deve sofrer modificações para se manter eficiente nessas condições [19]. Esquemas robustos de recuperação de erros podem ser empregados para minimização do efeito das altas taxas de perdas. Essa solução, porém, acaba por diminuir a eficiência de utilização do canal de transmissão já que dados redundantes serão transmitidos, prejudicando ainda mais a vazão.

Além das modificações necessárias a serem realizadas na camada transporte, outras camadas da arquitetura de protocolos TCP/IP necessitam de mudanças. O IP, o protocolo de camada rede da Internet, por exemplo, precisa passar a dar suporte ao nó móvel, permitindo que pacotes destinados a ele sejam entregues mesmo que ele não se encontre em sua rede de origem. Extensões ao IP vêm sendo propostas para lidar com a mobilidade [20, 21, 22].

Existem ainda outras questões típicas de ambientes móveis, onde há a utilização de equipamentos portáteis. Algumas dificuldades são relacionadas, por exemplo, com a durabilidade das baterias. Tais questões podem aparentar não ter relação direta com a transmissão de dados, entretanto, até mesmo a qualidade de serviço obtida nas transmissões pode ser influenciada por políticas de economia de energia. Conseqüentemente, tais políticas poderão influenciar na carga exercida por um terminal sobre a infra-estrutura de rede. Uma menor potência despendida com as transmissões iria prolongar o uso da bateria, reduzindo entretanto a taxa máxima de transmissão [13]. A qualidade de exibição das mídias pode variar para permitir uma otimização do tempo de utilização da bateria. A opção por uma qualidade de exibição mais baixa implicaria novamente em uma carga menor na rede.

Além dos problemas relacionados à eficiência das transmissões, outras características, como uma maior dificuldade para garantia de segurança, são inerentes à redes móveis. Essas questões vão desde a transmissão de dados, que é realizada via rádio podendo ser interceptada com facilidade, até aspectos de autenticação de usuário, uma vez que equipamentos móveis são mais suscetíveis a perdas e roubos [23].

2.2.1 *Handover*

Um dos possíveis cenários de utilização das redes sem fio é obtido tratando-as como redes de acesso à uma infra-estrutura fixa, como a Internet. Nesse caso um provedor de acesso possui uma rede sem fio com pontos de acesso conectados a um roteador de borda de uma infra-estrutura IP [24]. Dessa maneira, diferentes provedores de acesso móvel podem se conectar à rede fixa. Os usuários de terminais móveis têm então a possibilidade de se mover dentro da rede sem fio de seu provedor assim como para os demais provedores de acesso, através de esquemas de *roaming*. As redes sem fio podem envolver diferentes tecnologias. Entre outras, elas podem ser redes locais sem fio adequadas para uso dentro de prédios, redes celulares sem fio para regiões metropolitanas ou redes utilizando satélites para áreas mais amplas. Tipicamente, conforme apresentado na figura 2.1, as redes de acesso sem fio são interligadas por uma infra-estrutura fixa de acesso, uma rede de núcleo. Esta conecta diferentes domínios/redes de acesso sem fio, faz o transporte entre essas redes de acesso e permite a interconexão com redes fixas externas.

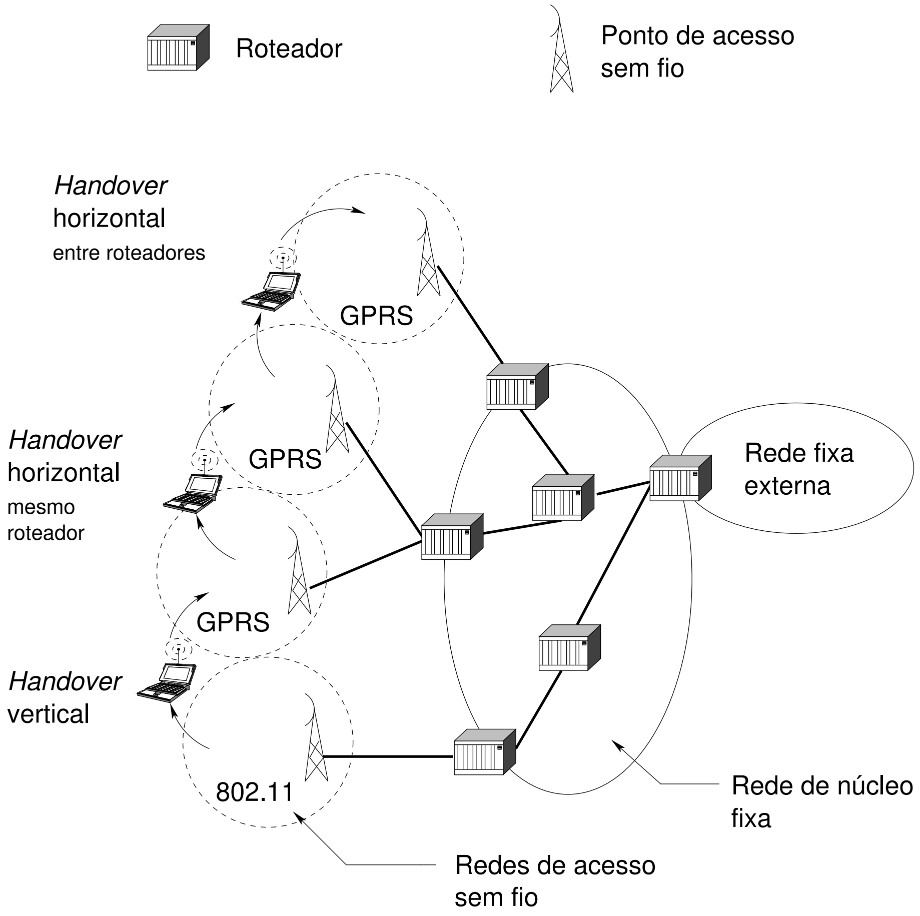


Figura 2.1: Arquitetura típica.

De acordo com a figura 2.1, observa-se a possibilidade de um nó, ao se mover, passar de um domínio de acesso para outro adjacente durante a transmissão. Essas mudanças são denominadas de *handovers*. Ao se passar de um domínio que utiliza uma determinada tecnologia de transmissão sem fio para domínios que utilizem outro tipo de tecnologia, como o que ocorre no caso da saída do interior de um prédio (ocorrendo a modificação da conexão de uma rede local sem fio para uma rede metropolitana sem fio), é dito que se realiza um *handover* vertical. No caso das tecnologias serem equivalentes, como no caso de uma rede celular, onde existem várias estações rádio base de alcance limitado possuindo mesma tecnologia de transmissão, os *handovers* são ditos horizontais [25]. Nos *handovers* verticais podem ocorrer mudanças abruptas em características do enlace devido às mudanças de tecnologia.

No caso de um nó móvel sair de um prédio para ambiente externo, ele poderia estar trocando a conexão com uma rede local sem fio como uma rede baseada no padrão IEEE 802.11 de 11 Mbps por uma rede metropolitana baseada em GPRS [26] de 160 kbps. Dessa forma, nota-se que as variações podem ser muito grandes, além de freqüentes [6].

Nos casos de *handover* vertical, as variações de recursos disponíveis, como banda passante, são claras. Mesmo nos casos de *handover* horizontal grandes variações podem ocorrer com freqüência. Nesses casos, as variações do serviço ocorrem por causa de possíveis sobrecargas em pontos do sistema. Conforme foi colocado inicialmente, as redes móveis são redes de acesso conectadas através de um roteador de borda a uma infra-estrutura de rede fixa. Já que os nós são móveis, podem ocorrer situações onde um grande número de nós se encaminhe para uma determinada rede de acesso, gerando uma sobrecarga em seu roteador de borda ou em determinados enlaces no seu interior. Essa incerteza com relação ao volume de tráfego é diretamente influenciada pela mobilidade [24, 6], trazendo maior complexidade à tarefa de realizar um provisionamento que garanta a qualidade de serviço desejada aos usuários.

A arquitetura descrita na figura 2.1 exemplifica ainda situações em que a realização de um *handover* desencadeia ou não o estabelecimento de uma diversidade de novos caminhos na rede de núcleo. No caso mais simples, apenas o ponto de acesso sem fio é trocado. Em outras situações, pode ser necessária a modificação de caminhos na rede, ou até mesmo uma troca de roteador utilizado para conexão com a rede de núcleo. Os autores de [27] classificam essa diversidade de utilização de novos caminhos em diferentes "profundidades de *handover*". A realização do *handover* em cada uma das diferentes "profundidades" traz dificuldades específicas ao sistema móvel. No caso de mudança somente de ponto de acesso, questões relacionadas à disponibilidade de canais de rádio devem ser consideradas. Alguns canais podem ser reservados para nós em movimento que cheguem à região, de forma a minimizar a probabilidade de bloqueio dessas chamadas [6, 14]. Já os *handovers* mais "profundos", que envolvam a troca de roteadores na rede de núcleo, trazem desafios aos esquemas de provisionamento de recursos, para que se mantenham a qualidade de serviço oferecida nessas situações de mobilidade. Técnicas de predição de movimento podem ser aplicadas para que se reservem recursos em ambas as redes, tanto nas de acesso sem fio como na de núcleo.

A classificação dos *handovers* pode ainda ser baseada na realização ou não de uma transição abrupta da transmissão durante uma migração entre regiões. Assim podem ocorrer o *hard handover* ou o *soft handover* [28]. O *hard handover* ocorre em sistemas de telefonia celular analógicos ou que utilizem acesso de rádio TDMA, por exemplo. Além desses casos, o *hard handover* ocorre quando o nó se move entre redes de acesso que não possuam conexão direta, sendo necessário que se troque de roteador de borda na rede de núcleo. Nesse processo, o tempo despendido para restabelecimento da conexão após a troca de região é maior do que no *soft handover* (geralmente 100 ms ou mais), podendo ocorrer pequenas interrupções no serviço, perceptíveis ao usuário. As chances de perda da conexão nesse caso são superiores àquelas do *soft handover*.

O *soft handover*, que é a transição suave entre regiões, é realizado em sistemas baseados em acesso de rádio CDMA. Nesse processo, o nó móvel é transferido entre as regiões dentro de uma mesma rede de acesso sem fio de forma mais suave. Esse *handover* ocorre de forma muito rápida (tipicamente na ordem de 20 ms) e tem uma baixa probabilidade de desconexão. Para realizar um *soft handover* o sistema deve permitir que o terminal móvel se comunique simultaneamente com mais de um ponto de acesso sem fio. Assim, ao se mover na fronteira entre duas regiões, o nó móvel se conecta com os dois pontos de acesso. Quando o sinal de rádio de um dos pontos de acesso se tornar insuficiente, o equipamento móvel descarta essa conexão continuando a transmitir dados através do ponto de acesso com sinal mais forte. Essa característica é conhecida como macro-diversidade. Além de permitir que o *handover* ocorra de forma suave, a macro-diversidade faz com que a transmissão fique mais robusta, levando a uma redução nas taxas de erro, uma vez que os sinais recebidos dos diferentes pontos de acesso podem ser combinados em um mais confiável.

A necessidade de manter mais de uma conexão referente a um único fluxo de usuário dentro da rede de acesso traz uma maior rigidez com relação à necessidades de qualidade de serviço. A rede de acesso precisa assegurar um alto sincronismo na chegada de quadros para que as diferentes conexões possam ser combinadas em um único fluxo.

Como foi exposto ao longo dessa seção, os sistemas móveis estão sujeitos a grandes variações nos parâmetros de rede, como vazão disponível, sendo clara a necessidade de esquemas adaptativos para a transmissão de dados em tais ambientes. Dentre as diversas particularidades das redes móveis, os *handovers* merecem uma atenção particular, por serem conseqüência direta do movimento dos nós. Eles são os responsáveis pelas maiores variações nos recursos disponíveis para cada usuário. Além disso, diferentemente das variações que sofrem influência direta do meio ambiente como as variações no canal de comunicação, as variações causadas pelos *handovers* podem ser controladas de forma a otimizar a relação entre QoS e utilização do sistema.

2.3 Redes celulares

A popularidade das comunicações móveis tomou grande impulso a partir da criação das redes de telefonia celular, que vieram a solucionar os problemas dos esquemas precursores de telefonia móvel, como capacidade muito limitada de oferta de serviços, baixa qualidade dos mesmos e a utilização ineficiente do espectro de freqüência (cada faixa de freqüência podia servir somente a um usuário, a cada instante, em uma ampla área geográfica).

Historicamente, o crescimento das comunicações móveis foi um pouco lento, até que durante os anos 50 e 60 a AT&T Bell Laboratories e outras empresas de telecomunicações desenvolveram o conceito de célula. Este conceito, aliado ao desenvolvimento de novas tecnologias, repercutiu em grande investimento por parte dos fabricantes levando ao surgimento dos vários sistemas de telefonia celular.

As redes celulares vêm então sofrendo transformações significativas desde o início da década de 90, com a passagem da tecnologia analógica (primeira geração, sendo o sistema AMPS um exemplo) para a digital (segunda geração, representados pelo GSM e IS-95 por exemplo) [13]. Os sistemas de primeira e segunda geração representam etapas na evolução da comunicação móvel em direção a um sistema que opere globalmente na chamada terceira geração (*3rd Generation* - 3G).

O conceito de célula foi um dos principais avanços que permitiu aumentar-se consideravelmente o número de usuários dividindo simultaneamente um espectro de freqüência limitado[2]. Este conceito consiste na divisão da área de cobertura do sistema de telefonia móvel em áreas menores denominadas células, possibilitando assim a utilização de transmissores de baixa potência e uma otimização do espectro através da reutilização das freqüências.

A estrutura básica de um sistema celular, conforme apresentado na figura 2.2, é composta basicamente por um terminal móvel; uma estação rádio-base que funciona como ponto de acesso dos móveis para a infra-estrutura fixa, estabelecendo o enlace de rádio para os diversos terminais móveis que se encontram dentro da sua área de cobertura; uma rede de acesso, para interligar as diferentes estações rádio base; e uma rede de núcleo que interliga diferentes redes de acesso entre si e com o restante do sistema de telefonia fixa.

As primeiras redes celulares foram idealizadas para oferecer um serviço de telefonia móvel ao seus usuários. Assim, as redes celulares de primeira e segunda geração são otimizadas para o transporte de voz, oferecendo serviços de mobilidade com conexões de boa qualidade por serem construídas em redes de comutação de circuitos. Essa infra-estrutura apresenta falta de flexibilidade, em oposição às redes orientadas a pacotes. Além disso, a utilização de circuitos dedicados resulta em uma elevação dos custos do serviço para o usuário final. Isso ocorre porque no caso do transporte de mídias que apresentam comportamento de transmissão em rajadas, como ocorre na transmissão de dados, a utilização de um circuito dedicado pode representar uma grande sub-utilização dos recursos de rede. Assim, os sistemas de segunda geração não visavam inicialmente o transporte de outras mídias além da voz, como dados e vídeo, fazendo com que o transporte dessas acabasse por se tornar ineficiente e caro.

[2] Além do conceito de célula, fatores como o desenvolvimento de novos CODECs, o surgimento de mecanismos de controle de potência, a evolução das técnicas de acesso ao meio, entre outros, foram importantes para o aumento da capacidade dos sistemas.

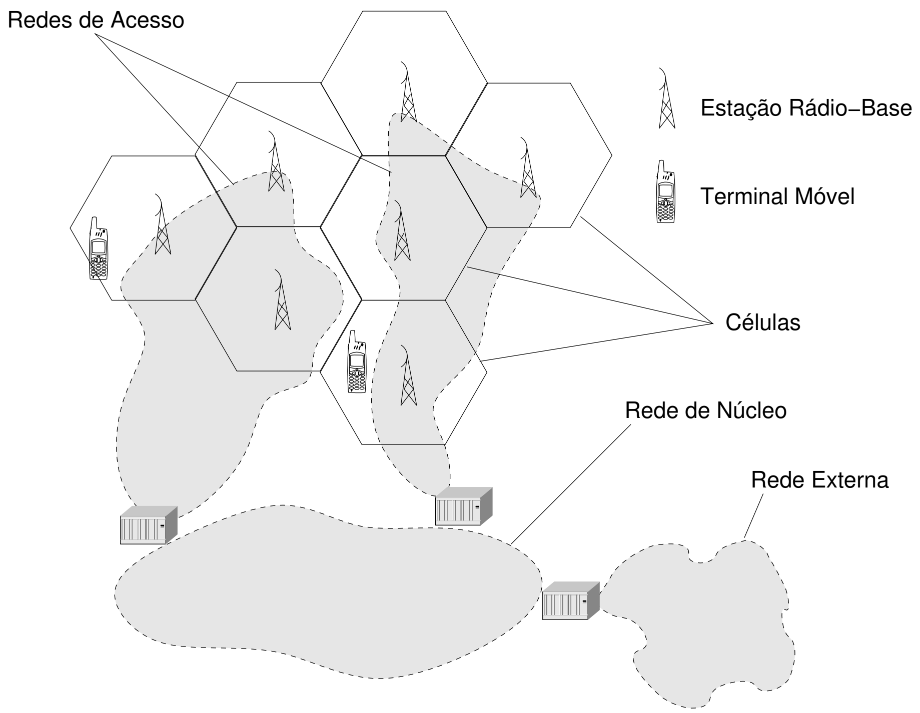


Figura 2.2: Arquitetura de um sistema celular.

As redes móveis 3G irão substituir o paradigma de comutação de circuitos pela comutação de pacotes [24]. Essa transição vem sendo realizada na chamada geração 2,5[3] de telefonia celular, onde o tráfego de dados utiliza uma infra-estrutura de comutação de pacotes que coexiste com a estrutura de comutação de circuitos[4]. Entre as vantagens oferecidas pela comutação de pacotes, quando comparada à comutação de circuitos, encontram-se o melhor aproveitamento do meio de transmissão, por causa de ganhos estatísticos, e uma melhor integração entre serviços de transmissão de dados e multimídia [13, 29].

O maior exemplo de sucesso de serviço baseado em comutação de pacotes é a Internet, que já tem papel fundamental nas telecomunicações. Sua importância como uma solução global de telecomunicações deve ser considerada para o futuro. Assim, as soluções para transmissão multimídia e de dados em um ambiente de redes de acesso sem fio devem permitir a interconexão transparente com o restante da Inter-

[3] A geração 2,5G é uma evolução intermediária entre 2G e 3G.

[4] A tecnologia GPRS empregada em conjunto com o GSM (*Global System for Mobile Communications*) opera dessa forma [26].

net. Portanto, quando se buscam arquiteturas para acesso móvel, é interessante que se utilizem redes IP, uma vez que toda a Internet é baseada nessa tecnologia. O IP traz as vantagens da comutação por pacotes e possui ainda soluções consolidadas para as principais dificuldades desse tipo de transmissão. Os esquemas de endereçamento, roteamento, qualidade de serviço, entre outros, são amplamente estudados e conhecidos nessas redes. A utilização da tecnologia IP possui, além da flexibilidade, robustez e propriedade escalar [30].

A utilização do protocolo IP também permite uma padronização em nível de aplicações. O projeto de aplicações e serviços pode então ser realizado de forma independente ao ambiente de utilização, seja ele fixo ou móvel. Dessa forma, até mesmo as aplicações já utilizadas nas redes fixas baseadas em IP podem ser aproveitadas para ambientes de mobilidade, desde que o IP modificado para lidar com a mobilidade consiga torná-la transparente às aplicações [27, 31].

2.3.1 Sistemas celulares de terceira geração

Os sistemas de comunicações móveis 3G têm como objetivo permitir uma integração completa e transparente entre os ambientes fixos e móveis, permitindo acesso ubíquo às informações. Para facilitar essa integração, seguindo os argumentos expostos na seção anterior, existe uma tendência de utilização da tecnologia IP fim-a-fim nesses sistemas [32, 33, 34, 35]. Uma das principais vantagens dessa opção é a interconexão das redes celulares com a Internet de forma transparente.

Os sistemas 3G devem oferecer uma conexão ininterrupta, de alcance global, permitindo uma comunicação pessoal que ofereça toda uma nova gama de serviços, como transmissões multimídia e acesso à transmissão de dados. Para isso, os sistemas 3G devem ser capazes de lidar com os mais diversos tipos de aplicação, dando suporte a diferentes classes de tráfego, desde voz em tempo real até tráfego de dados em rajadas.

Para fornecer um suporte adequado às aplicações, as especificações dos sistemas 3G incluem a oferta de altas taxas de dados (384 Kbps em ambientes de alta mobilidade e até 2 Mbps em ambientes mais estáticos), utilização de tecnologia de comutação de pacotes e oferta de qualidade aos serviços comparável àquela atualmente disponível. Além disso, os sistemas 3G devem ser capazes de coexistir com diferentes sistemas de comunicação móvel, como as redes celulares 2G e redes de satélite atualmente existentes [14, 24].

Um dos sistemas celulares que mais vem se destacando como padrão para a terceira geração é o sistema *Universal Mobile Telecommunications System* (UMTS) [34], que será apresentado na próxima seção.

Elementos da arquitetura UMTS

A arquitetura de rede celular de terceira geração definido pela *Third Generation Partnership Project* (3GPP) [36] é o sistema *Universal Mobile Telecommunications System* (UMTS) que possui capacidade de comutação de pacotes desde sua interface aérea até o roteador de saída de sua rede de núcleo. O UMTS é uma evolução do sistema *General Packet Radio Service* (GPRS) [37, 38] de 2,5G, que por sua vez é uma melhoria implementada na infra-estrutura GSM (*Global System for Mobile Communications*) para prover um serviço orientado a pacotes aos usuários. Sua arquitetura, descrita de forma simplificada na figura 2.3, é composta por redes de acesso (*UMTS Terrestrial Radio Access Network* - UTRAN) e uma Rede de Núcleo (*Core Network* - CN).

A UTRAN é composta de um ou mais sistemas de rede de rádio (*Radio Network System* - RNS) que inclui estações rádio base (*Base Station* - BS) e controladores de rede de rádio (*Radio Network Controllers* - RNC). Dessa forma uma UTRAN tem capacidade de servir diversas BSs, cada uma delas atendendo a muitos terminais móveis (*Mobile Station* - MS). Conforme descrito na seção 2.2.1, os MSs podem se conectar com mais de uma BS adjacente, de forma a viabilizar a realização de *soft handovers*.

A Rede de Núcleo oferece transporte entre dois nós de borda permitindo interconectividade entre diversas UTRANs. Ela permite também a comunicação com

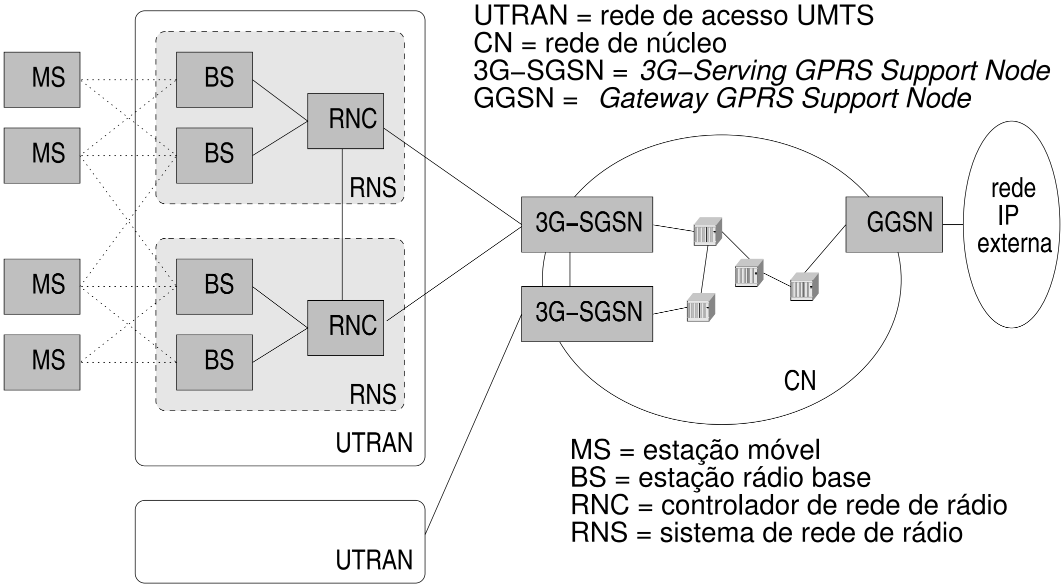


Figura 2.3: A arquitetura UMTS.

redes externas de outros operadores, inclusive com redes IP como a Internet. Os elementos de borda da Rede de Núcleo são os 3G-SGSN (*3G-Serving GPRS Support Node*) e o GGSN (*Gateway GPRS Support Node*). O GGSN é o portal entre a Rede de Núcleo e a rede de pacotes externa. Esse nó é visto pela rede externa como um roteador simples, servindo os endereços dos nós móveis. O 3G-SGSN é a interface entre a Rede de Núcleo e a UTRAN, realizando o roteamento dos pacotes que chegam endereçados às MSs para o RNS adequado (aquele que está servindo a MS no momento) [26, 37]. É importante notar que um único 3G-SGSN pode servir múltiplos RNSs. Assim, mesmo que cada terminal móvel efetue transmissões a baixas taxas, a taxa total agregada nos 3G-SGSN, e em conseqüência nos enlaces do interior da Rede de Núcleo, poderá ser bastante elevada [39].

No cenário da figura 2.3, um operador de rede celular possui nós 3G-SGSN servindo diferentes áreas. Assim os nós móveis são capazes de se mover dentro da área controlada por um 3G-SGSN assim como entre áreas servidas por diferentes 3G-SGSN.

O número de *handovers* efetuado durante uma chamada/transmissão do usuário móvel nas redes 3G deve ser maior do que o observado para as redes atuais[5], por causa das novas aplicações que podem exigir transmissões durante um tempo prolongado. Além disso, existe uma tendência de diminuição no tamanho das células nas áreas de alta utilização do sistema, acarretando em um aumento no número de células. Isso é necessário para que se possa aumentar a capacidade de transmissão. Esse maior número de células também tende a aumentar número de *handovers* que ocorrem durante cada chamada. Conseqüentemente, os *handovers* que envolvem trocas de pontos de acesso da rede de núcleo também vão ter uma maior chance de ocorrer. Assim, as dificuldades trazidas com os *handovers* terão uma maior influência nas redes 3G, quando comparadas às redes móveis atuais, trazendo novos desafios à provisão adequada de QoS.

QoS em UMTS

Os sistemas celulares de terceira geração devem dar suporte à diferentes tipos de mídias, respeitando sempre os requisitos de QoS para cada tipo de aplicação, incluindo serviços de tempo real com garantias. Esse tipo de tráfego deve representar uma porção significativa do total transportado nas redes 3G, uma vez que elas são uma evolução dos atuais sistemas celulares 2G, voltadas principalmente para telefonia (serviço de voz interativo). Assim, os tradicionais serviços de telefonia devem se somar às aplicações de tele-conferência e videofonia, mantendo grande parte do tráfego das próximas gerações de comunicação pessoais com requisitos de transmissão em tempo real.

Para tanto, tais sistemas devem permitir um provisionamento de QoS fim-a-fim. Assim, a arquitetura de QoS do UMTS é baseada em um serviço fim-a-fim, que entretanto é decomposto em partes. Esse é o serviço baseado em diferentes serviços de portador (*bearer services*), como apresentado na figura 2.4.

Na arquitetura UMTS, um serviço de portador define as características de QoS entre pontos-finais da conexão, sendo que o serviço fim-a-fim é visto como uma composição de diferentes serviços oferecidos em cada uma das redes que compõem

[5] As redes GSM geram uma média de um a dois *handovers* por chamada [40].

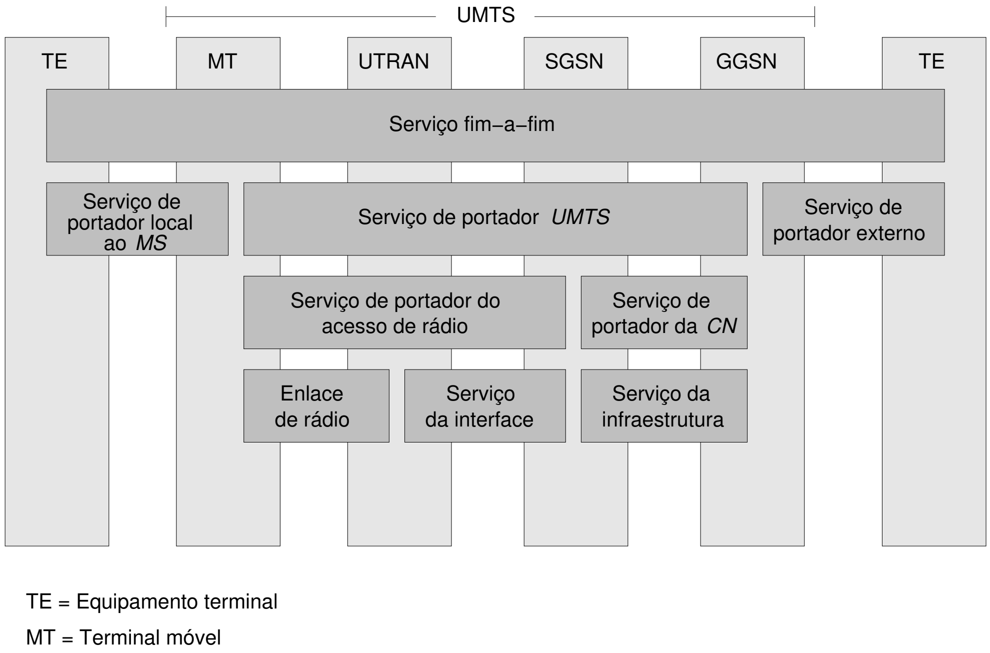


Figura 2.4: A estrutura dos serviços de portador para o UMTS.

o sistema. A QoS oferecida por cada uma dessas redes é fundamental para a composição da QoS fim-a-fim obtida. Dessa forma, observa-se que a estrutura de QoS é organizada em camadas e o serviço fim-a-fim é subdividido em [41, 42]:

1. serviço de portador local ao sistema móvel (*mobile system local bearer service*) - é o responsável pela QoS nos equipamentos do usuário, como *laptops* e telefones móveis. Cada equipamento do usuário utiliza um portador local (*local bearer*), como por exemplo uma interface de programação orientada a QoS;

2. serviço de portador UMTS (*UMTS bearer service*) - é o serviço de QoS oferecido pelo operador do sistema 3G. Ele deve assegurar a QoS dentro do sistema e realizar as operações necessárias para inter-operar com as redes externas;

3. serviço de portador externo (*external bearer service*) - é o suporte a QoS oferecido pelas redes externas, como por exemplo a Internet.

O serviço de portador UMTS, que é o responsável pela QoS dentro do domínio do operador 3G, é sub-dividido em serviço de portador do acesso de rádio (*radio access bearer service*) e serviço de portador da Redes de Núcleo (*core network bearer* *service*). O primeiro é o serviço de QoS oferecido pela UTRAN, que envolve a transmissão nos enlaces de rádio e em toda a rede de acesso, desde o equipamento móvel até o nó de borda da Rede de Núcleo, o 3G-SGSN. O segundo é responsável pela QoS dentro da Rede de Núcleo, provendo garantias de transporte de dados entre um 3G-SGSN e um GGSN.

A questão de provisionamento de QoS na UTRAN é bastante delicada, envolvendo características da transmissão de rádio. Ao receber um pacote de dados, a UTRAN realiza diversas tarefas relacionadas com a otimização da transmissão dos mesmos através do enlace de rádio [28, 33]. Além disso, questões relacionadas com a macro-diversidade (seção 2.2.1) aumentam a necessidade de sincronismo forçando baixos atrasos nessas redes de acesso. Assim, os requisitos QoS de todo o tráfego nas UTRANs, particularmente de atraso, são muito estreitos. Alguns exemplos numéricos podem ser encontrados em [33]. A QoS nessas redes deve então ser garantida de forma rígida. Devido à maturidade do suporte de QoS nas redes ATM, esta é a tecnologia de transporte mais empregada para as UTRANs.

No caso de uma rede totalmente baseada em tecnologia IP, os esquemas de QoS propostos pela *Internet Engineering Task Force* (IETF) [5] em [3, 2] podem ser utilizados para assegurar a qualidade necessária, tanto nas UTRANs como também na Rede de Núcleo. Para que se obtenham as garantias rígidas necessárias nas UTRANs será necessário que se empregue técnicas bastante conservadoras de controle de tráfego na mesma. Já na Rede de Núcleo, que possui requisitos mais relaxados de QoS, podem ser oferecidas classes de serviço com qualidade diferenciada, respeitando as características mais importantes de cada aplicação e buscando manter um alto índice de ocupação dos recursos da rede, como será discutido na seção 4.1.1.

Um aspecto importante que deve ser levado em conta quando se buscam soluções que possam oferecer QoS nas Rede de Núcleo, é a capacidade de tais soluções em lidar com uma grande escala de fluxos de usuários. As redes de acesso 3G podem consistir em centenas de RNCs, estando cada uma delas conectada a diversas BSs. Assim, o volume de tráfego que passa através de cada RNS pode variar de uns poucos até milhares de fluxos simultâneos [39]. Dessa forma, os roteadores da Rede de Núcleo deverão ser capazes de lidar com alguns milhares de fluxos de usuários de forma simultânea. Esse número pode ainda sofrer grandes variações, exigindo uma capacidade de adaptação do provisionamento de QoS, de forma a se evitar tanto os desperdícios de recursos como o não atendimento dos requisitos de QoS das diversas aplicações. Para que se possa oferecer QoS em redes IP com propriedade escalar, os trabalhos apresentados em [41, 37] sugerem a aplicação de uma das propostas de QoS do IETF, a diferenciação de serviços (*Differentiated Services* - DiffServ [3]), na Rede de Núcleo. A escolha do DiffServ para dar o suporte de QoS é também motivada pelo fato dessa tecnologia ser a mais promissora para ser utilizada pela rede IP externa (serviço de portador da rede externa - *external bearer service*), no caso a Internet.

O DiffServ, entretanto, não foi planejado para trabalhar em ambientes de grande variação de volumes de tráfego, como pode ocorrer no caso dos ambientes de mobilidade. Nesses ambientes, a utilização do DiffServ pode levar a violações de QoS. Para evitar essas violações, o DiffServ passa a depender de um provisionamento conservador, que pode resultar em grandes desperdícios de recursos. Uma discussão mais completa sobre os possíveis esquemas de QoS que podem ser aplicados nas redes de núcleo de ambientes de mobilidade será realizada no capítulo 3.

2.4 QoS e mobilidade

Conforme apresentado na seção 2.3.1, há uma forte tendência para que se utilize o protocolo IP como base da arquitetura de redes nos sistemas de terceira geração. É necessário, portanto, que o IP seja capaz de transportar diferentes tipos de tráfego, incluindo tempo real, para que se possam fornecer os serviços planejados para o sistemas 3G. Uma grande dificuldade a ser solucionada no contexto de transporte de tráfego em tempo real, como voz e vídeo interativos, é o provisionamento adequado de QoS, particularmente em redes móveis. O movimento dos nós dificulta o planejamento adequado da rede, dificultando a oferta desses serviços. Em ambientes de mobilidade, para que a QoS oferecida não se torne inadequada, a necessidade de mecanismos eficientes de controle de tráfego torna-se ainda maior.

O desenvolvimento inicial do protocolo IP não tinha como objetivo o transporte de tráfego com requisitos temporais. A Internet era originalmente utilizada para o transporte de arquivos e eventualmente para acesso remoto de servidores. Tais aplicações são bastante flexíveis quanto à qualidade de serviço obtida. Já o tráfego multimídia, incluindo transporte de voz e vídeo em tempo real exige que limites mínimos de QoS, como atraso, variação no atraso (*jitter*) e taxa de perdas sejam garantidos.

As propostas de QoS para redes IP, particularmente para a Internet, vêm sendo amplamente estudadas e discutidas no âmbito da IETF. Tais propostas serão apresentadas no capítulo 3.

Em uma arquitetura de QoS é importante que se procure respeitar as características de tráfego das diferentes aplicações. Os requisitos importantes para um tipo de aplicação podem não ser fundamentais para outras. Assim, os parâmetros de QoS menos importantes, aqueles que influenciam menos o desempenho percebido por um dado usuário, podem ser exploradas. Nesse caso, podem ser arquitetados serviços que realizem uma degradação em alguns parâmetros, quando houver uma sobrecarga do sistema, para otimizar o atendimento de uma forma global. Isso permitiria que um maior número de usuários se mantivesse satisfeito, mesmo que uma pequena queda na QoS fosse necessária. Nesse momento deve ser identificado o que é mais importante para cada tipo de aplicação: a confiabilidade, a disponibilidade/conectividade do serviço *etc.*. Deve ser lembrado que as aplicações voltadas para mobilidade devem ser capazes de se adaptar, até certos limites, às condições de QoS oferecidas pela rede, uma vez que essa QoS pode apresentar variações.

Os parâmetros de QoS normalmente abordados em redes fixas tradicionais são atraso, variação no atraso (*delay jitter*), vazão e taxa de perdas. As aplicações são classificadas de acordo com as suas necessidades específicas, baseando-se nesses parâmetros. A maioria das aplicações é extremamente sensível a alguns desses parâmetros de QoS, mas por outro lado há quase sempre uma flexibilidade com relação aos parâmetros restantes. O transporte de dados, por exemplo, é geralmente diretamente influenciado pela taxa de perdas no meio, apresentando uma maior tolerância ao atraso. Tráfegos de tempo real, como voz, se comportam de forma oposta, sendo altamente sensíveis à atrasos e relativamente robustos a perdas. De uma forma geral o atraso é importante para aplicações que necessitam de interatividade. Os fluxos que representam mídias contínuas (voz e vídeo) a serem reproduzidos no receptor (os chamados *streams* de dados) necessitam de pequenos valores de *jitter*, de forma a reduzir o tamanho dos *buffers* de reprodução da mídia [43]. A taxa de perdas deve ser mantida baixa nas aplicações onde a confiabilidade na transmissão é fundamental. Assim, as retransmissões podem ser evitadas e o desempenho do sistema é elevado. A alta vazão é significativa quando se realizam transferências de grande volume de dados, assim como para o transporte de mídias que demandam alta utilização dos enlaces.

Os parâmetros de QoS descritos acima também devem ser observados nos ambientes de mobilidade. Entretanto, o fato da mobilidade dos usuários não ser previsível exige que novos parâmetros de QoS sejam definidos. A garantia de serviços sem interrupções é um deles. Em muitos casos, o fato de se estar continuamente conectado pode ser mais importante do que a ausência total de perdas ou até mesmo do que a existência de uma alta vazão. Isto é verdade para tráfegos de voz interativos por exemplo. Um usuário pode preferir uma diminuição na qualidade da conversação em vez de ter sua chamada interrompida de forma abrupta. Nesse exemplo, fica clara também outra característica da arquitetura de QoS que deve ser explorada, a degradação suave do serviço. A degradação suave pode ser utilizada para assegurar uma baixa probabilidade de desconexão. Neste contexto aparecem como parâmetros de QoS a probabilidade de conexão contínua e o perfil de perdas [6, 7].

É importante que se observe que o conceito de QoS para redes móveis deve ser mais flexível do que o existente em redes fixas [44]. De acordo com o que já foi apresentado nas seções 2.2 e 2.2.1, o movimento dos nós pode levar a diversas situações onde não é possível se manter as condições de QoS estáveis. Dessa forma, as aplicações voltadas para ambientes de mobilidade devem ser adaptativas, estando preparadas para lidar com níveis de variação no serviço oferecido [23]. Assim as características e a flexibilidade de cada aplicação devem ser exploradas de forma a otimizar globalmente o serviço oferecido.

Tabela 2.1: Satisfação do usuário para níveis de taxa de erro na transmissão de voz.

Taxa de perdas de pacotes (%)	MOS
0	5
10	4
20	3
30	2,5
40	2,5
50	1

Tomando-se como exemplo o tráfego de voz interativo, nota-se que o parâmetro de QoS fundamental é o baixo atraso. Um alto atraso faz com que os interlocutores percam a interatividade, o que torna a conversação inviável. É também fundamental que a conexão se mantenha durante toda a conversação, com probabilidades de desconexão mínimas. Para que se atendam às duas condições anteriores, consideradas fundamentais, pode ser necessário que haja uma degradação nos níveis de taxa de perdas. Apesar de ser muito importante que se mantenha a taxa de perdas baixa, existe uma certa flexibilidade com relação a esse parâmetro. A tabela 2.1 demonstra os níveis de satisfação do usuário, em uma escala de 1 a 5, para uma dada taxa de perdas, quando se utiliza o CODEC DVI4 em conjunto com esquemas de redundância [45]. Podemos observar que a satisfação do usuário não cai de forma abrupta, sendo possível que se explore melhorias na probabilidade de desconexão e na manutenção de baixos atrasos em troca de aumentos controlados na taxa de perdas.

Nesse exemplo, o usuário aceitaria certa flexibilização quanto aos requisitos de taxa de perdas para um tráfego de voz interativo em troca de uma maior conectividade com garantias de baixos atrasos. Esquemas de degradação de parâmetros de QoS em troca de uma maior utilização nos casos de sobrecarga também foram propostos em [46, 24, 47], entre outros.

Assim, a partir dos parâmetros de QoS apresentados, é possível que se planeje classes de serviço voltadas para mobilidade, explorando as necessidades e flexibilidades inerentes de cada tipo de aplicação. Os autores de [46, 32] discutem a criação de classes mais flexíveis para ambientes de mobilidade. No trabalho aqui apresentado, a criação de classes voltadas para mobilidade é discutida na seção 4.1.1.

O presente capítulo procurou realizar um estudo amplo sobre as características das redes móveis, concentrando-se nas questões que representam desafios ao provisionamento de QoS. O próximo capítulo irá abordar questões de QoS em Redes de Núcleo de sistemas 3G. A arquitetura de QoS empregada nessa rede deve ser capaz de lidar com grandes quantidades de usuários, sem deixar de oferecer a qualidade necessária e buscando otimizar a utilização global do sistema.

Capítulo 3

QoS em Redes de Núcleo

NESTE capítulo são descritas as principais soluções de QoS que podem ser aplicadas em Redes de Núcleo de sistemas 3G e suas características são analisadas. Na Seção 3.1 são apresentadas as duas arquiteturas de QoS mais discutidas no âmbito de pesquisa na Internet: o IntServ e o DiffServ. É então realizada uma análise crítica dessas propostas na seção seguinte. São também apresentadas outras propostas que possuem características híbridas. Finalmente, as principais necessidades e objetivos de um gerenciamento de recursos voltado para o provisionamento de QoS em ambientes de mobilidade são abordados na seção 3.3.

3.1 QoS na Internet

Os sistemas móveis 3G visam a dar suporte ao tráfego multimídia em redes de comutação de pacotes, oferecendo garantias de QoS fim-a-fim. Conforme o discutido na seção 2.3.1, a Rede de Núcleo do UMTS, baseada em IP, interconecta diferentes redes de acesso de um ou mais operadores, representando uma rede de alto volume de tráfego com diferentes requisitos de QoS. Além disso, a Rede de Núcleo realiza a conexão do sistema celular com o restante da Internet. Todo o tráfego de dados entre os nós móveis e a Internet passará pela Rede de Núcleo. Dessa forma, o desempenho da Rede de Núcleo afetará diretamente a QoS observada pelos nós móveis.

A comunidade de pesquisa em redes vem buscando diferentes soluções para permitir o controle de QoS em redes IP. Como discutido no capítulo 2, é interessante que a arquitetura 3G utilize esquemas que permitam a interoperação transparente entre o sistema celular e o restante da Internet, que é a rede de dados mais representativa nos dias de hoje. Isso inclui toda a tecnologia de transporte, indo desde o transporte IP até o suporte à QoS que será oferecido. As principais soluções de QoS IP voltadas para a Internet vêm sendo desenvolvidas nos grupos de trabalho do *"Internet Engineering Task Force"* (IETF) [5]. O IETF é o principal orgão responsável pela discussão e padronização de protocolos e das soluções para a Internet. As duas propostas de QoS mais consolidadas para utilização na Internet são a "Integração de Serviços" (IntServ) [2] e a "Diferenciação de Serviços" (DiffServ) [3, 4], que serão apresentadas nas próximas seções.

3.1.1 Integração de Serviços

A arquitetura de "Integração de Serviços", o IntServ, define extensões para a estrutura IP de forma a prover QoS fim-a-fim para cada micro-fluxo de dados das aplicações. O esquema define três classes de serviço que podem ser utilizadas de acordo com as necessidades de cada aplicação. A classe de serviços com requisitos mais rígidos é a classe "Serviços Garantidos" (*Guaranteed Services* [48]), que oferece limites superiores de atraso por pacote, porções garantidas de banda passante e ausência de descartes nas filas dos roteadores. Em seguida, a classe de "Carga Controlada" (*Controlled Load* [49]) oferece um serviço mais flexível, garantindo que limites de medidas estatísticas (como por exemplo atraso médio) não serão violadas mais vezes do que no caso de uma rede sem carga. Finalmente, a classe de "Melhor Esforço" (*Best Effort*) deve ser utilizada por aplicações que não possuam requisitos de QoS.

Para que o modelo IntServ possa fornecer garantias quantitativas às classes de "Serviço Garantido" e de "Carga Controlada" é necessário que mecanismos explícitos de sinalização e de controle de admissão por nó sejam empregados [50]. Esses mecanismos levam informações sobre os novos fluxos que chegam para cada elemento de rede no interior do caminho de comunicação. Assim, os roteadores podem realizar reservas de recursos para os novos fluxos, de forma a garantir seus requisitos de QoS. Caso não existam recursos disponíveis suficientes em um dado roteador no caminho, ele notifica explicitamente à aplicação a falha no processo de reserva, negando o acesso do novo fluxo à rede. Apesar da arquitetura IntServ não exigir a utilização de um protocolo de sinalização específico, o "Protocolo de Reserva de Recursos" (*Resource reSerVation Protocol* - RSVP) [51, 52, 53] é o mais comumente empregado [54].

O RSVP é então o protocolo utilizado pelas aplicações para realizar pedidos de reservas de recursos da rede. A resposta da rede é a admissão ou rejeição explícita do pedido. Combinando-se a arquitetura IntServ com o protocolo RSVP, obtém-se um modelo onde cada fluxo têm suas necessidades sinalizadas para os elementos de rede através da utilização do RSVP. Os elementos de rede aplicam então procedimentos de controle de admissão baseados nessas necessidades. Além disso, o elemento de rede configura mecanismos de controle de tráfego para garantir os serviços estabelecidos para aquele fluxo isoladamente dos restantes. Dessa forma, a sinalização através do RSVP configura classificadores de pacotes por micro-fluxo nos roteadores IntServ ao longo do caminho de comunicação.

O IntServ é então uma solução totalmente dinâmica, que requer uma sinalização explícita por fluxo. Tal solução provê excelentes garantias de QoS, oferecendo suporte ao tráfego de tempo real com requisitos rígidos de atraso e taxa de perdas. A comunicação que ocorre entre os elementos da rede e a aplicação, permitem uma utilização ótima dos recursos da rede, atingindo a maior utilização que permita que todos os usuários admitidos tenham níveis de QoS assegurados. O IntServ, entretanto, possui um grave problema causado pela alta carga de sinalização e pela necessidade de se armazenar nos elementos de rede um estado por micro-fluxo de cada aplicação. Assim, essa solução não é adequada para redes que precisam lidar com um grande número de fluxos, como o caso das Redes de Núcleo de sistemas celulares 3G. Essa dificuldade de atender a propriedade escalar é a principal desvantagem do modelo IntServ.

3.1.2 Diferenciação de Serviços

Em oposição ao modelo IntServ, a arquitetura de "Diferenciação de Serviços", o DiffServ, não busca um controle de QoS baseado em necessidades de micro-fluxos individuais. O provisionamento de QoS é implementado em nível de agregações de fluxos. Isso é possível a partir da observação que os diferentes micro-fluxos das diversas aplicações podem ser classificados em poucas categorias (classes de tráfego), dependendo das características das aplicações refletidas nas suas necessidades de QoS. Assim as garantias de QoS não são oferecidas por fluxo, mas por agregados de tráfego. As especificações dos agregados e do serviço são então descritas por contratos, os "Acordos de Nível de Serviço" (*Service Level Agreements* - SLAs).

O DiffServ utiliza um campo do cabeçalho dos pacotes (marcando o chamado *DiffServ CodePoint* - DSCP) para diferenciar conjuntos de fluxos, tratando-os como um conjunto/agregado. No interior do domínio DiffServ os agregados de fluxos de cada classe são tratados e encaminhados nó a nó de acordo com comportamentos pré estabelecidos, os "Comportamentos por Nó" (*Per Hop Behavior* - PHB's). Um PHB é "o comportamento do encaminhamento de pacotes observável externamente, aplicado por um nó DiffServ a um agregado" [4]. Cada classe de serviço dentro de um domínio DiffServ deve estar adequadamente provisionada para que se possa oferecer as garantias de QoS a cada agregado. Para que tal provisionamento se mantenha adequado utilizam-se mecanismos de controle de tráfego nas bordas dos domínios, onde o volume de tráfego é menor[1]. Isso torna a dificuldade no atendimento de uma larga escala de fluxos menos grave. Tais mecanismos incluem classificadores, medidores, e policiadores/condicionadores. Os medidores de tráfego são utilizados para indicar se as especificações de tráfego, como por exemplo sua taxa média e rajada máxima estão dentro do perfil contratado. Os marcadores de tráfego identificam a qual classe pertencem os pacotes dos fluxos. De acordo com a medição dos tráfegos, policiadores/condicionadores de tráfego podem ser empregados para limitar o volume de dados injetados por cada classe no domínio DiffServ, descartando

[1] Caso o tráfego na borda esteja concentrado em um único roteador de entrada, este deve ter capacidade de processamento para lidar com esse total de carga.

ou remarcando o tráfego excedente para classes inferiores ou mesmo modificando características das rajadas, realizando assim um controle de tráfego/admissão estático por classe em nível de pacotes [55]. O local de atuação dos condicionadores é ilustrado na figura 3.1. O tratamento adequado e o isolamento de cada classe são obtidos através da utilização de escalonadores de pacotes implementados em cada roteador, no interior do domínio DiffServ.

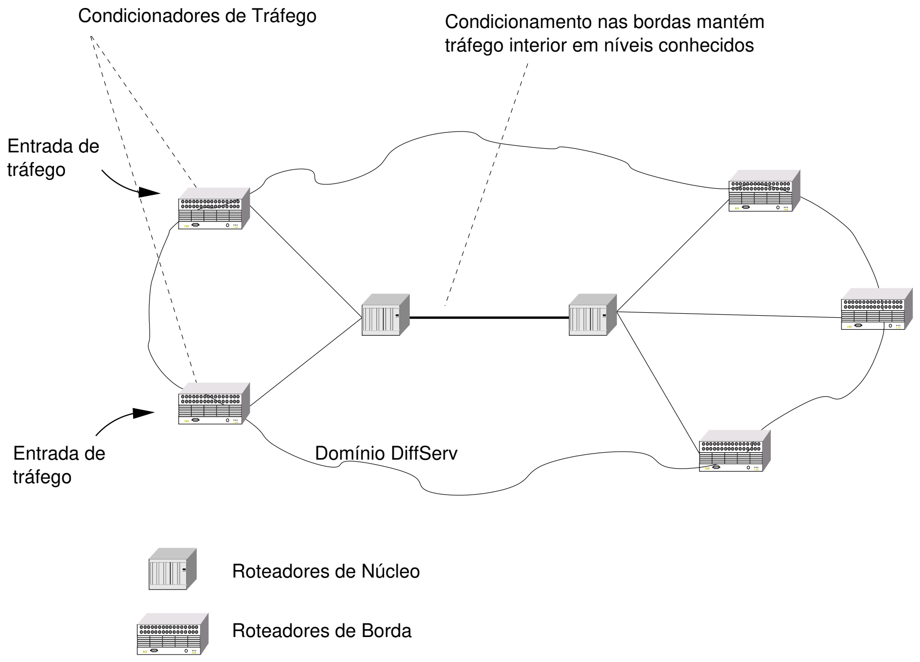


Figura 3.1: Condicionadores de tráfego no DiffServ.

A arquitetura DiffServ pode então ser vista como uma combinação: de marcação de pacotes nas bordas; utilização dessa marcação para diferenciar o tratamento em cada nó; e condicionamento de tráfego nas bordas, de acordo com suas marcações [27]. As regras de policiamento e condicionamento nas bordas do domínio em conjunto com os diferentes tratamentos dos pacotes nó a nó em seu interior (PHB) resultam em comportamentos por agregado em um domínio observáveis externamente, que foram denominados de "Comportamento por Domínio", (*Per-Domain Behavior* - PDB) [56].

Atualmente a arquitetura DiffServ oferece dois PHB's padronizados pelo IETF: O "Encaminhamento Expresso" (*Expedited Forwarding* - EF) [57, 58] e o "Encaminhamento Assegurado" (*Assured Forwarding* - AF) [59].

Através da utilização do PHB EF, pode ser obtido o serviço de "linha privativa virtual", que oferece garantias de baixo atraso, baixo *jitter* e pequena taxa de perdas, oferecendo uma banda passante assegurada ao agregado. Para realizar isso, é necessária a garantia de que a taxa de serviço nos nós do domínio seja superior à taxa de chegada de dados. Dessa forma, as filas nos roteadores estarão sempre minimizadas. Para tanto, esse serviço emprega condicionadores de tráfego nos nós de ingresso, de forma a garantir a correta relação entre as taxas. Assim, o EF pode ser utilizado para dar suporte a aplicações de tempo real.

O PHB AF, por sua vez, procura garantir vazão dos dados, permitindo a ocorrência de rajadas de pacotes, desde que estas sejam de curta duração. Desta forma os congestionamentos de curta duração são permitidos nas filas dos nós AF. Os congestionamentos longos, entretanto, são penalizados. Para isso, mecanismos de gerenciamento ativo de filas, como RED [60] são utilizados.

A arquitetura DiffServ oferece diferentes níveis de QoS sem exigir uma negociação explícita de cada aplicação para os fluxos que chegam na rede. Isso evita a sobrecarga gerada pela sinalização por micro-fluxo, assim como o atraso inicial de negociação para cada novo fluxo. O DiffServ se caracteriza como uma solução com ótimas propriedades para atender redes com larga escala de micro-fluxos, como o caso das Redes de Núcleo de sistemas celulares 3G. Isso ocorre por que a única operação realizada pelos roteadores no interior do domínio é a classificação de pacotes baseada em um campo do cabeçalho em um número limitado de classes e o tratamento diferenciado das mesmas. As funcionalidades mais complexas e que exigem o tratamento por micro-fluxo no DiffServ são realizadas nos roteadores das bordas do domínio. Dessa forma, o DiffServ pode ser visto como uma solução que diminui a complexidade dos roteadores de núcleo, onde o problema de atendimento de fluxos em grande escala é grande, trazendo-a para as bordas, onde o número de fluxos é reduzido.

Assim, o modelo DiffServ tem como base a correta configuração estática dos seus elementos (condicionadores e escalonadores) para que o atendimento às classes de serviço ocorra de forma adequada. Apesar da arquitetura do DiffServ limitar o volume e o padrão de tráfego na entrada da rede a valores conhecidos, ela não oferece meios para que se determine o comportamento desses fluxos no interior do domínio. Assim, para que se realize um provisionamento adequado da infra-estrutura de transmissão, é importante o conhecimento dos volumes e padrões de tráfego de cada classe atravessando cada nó do domínio [50].

A imprevisibilidade do movimento dos nós, entretanto, faz com que tal provisionamento não seja uma tarefa fácil. Conforme discutido na seção 2.2.1, o provisionamento estático pode tornar-se inadequado quando os nós móveis se concentram em um determinado roteador de borda, congestionando enlaces no domínio e degradando os níveis de QoS.

Cada nova configuração dos nós móveis pode fazer com que os fluxos que entram em cada roteador de ingresso fluam para qualquer um dos roteadores de egresso da região DiffServ. Em um caso extremo, uma porção substancial do tráfego pode se concentrar em determinados nós. Eventuais mudanças no roteamento dos pacotes podem também sobrecarregar alguns nós da região DiffServ, degradando a QoS obtida. Essas situações são mais comuns nos ambientes de mobilidade do que nas redes fixas tradicionais. Uma forma indesejável de se evitar as situações de sobrecarga seria a realização de um provisionamento acima das necessidades, um ineficiente super-provisionamento.

Em redes fixas tradicionais, provedores de serviço poderiam evitar as sobrecargas sem abrir mão de uma configuração eficiente através de restrições na generalidade do serviço prestado. Um exemplo seria a disponibilidade de serviços diferenciados para um agregado de fluxos desde que ele estivesse associado a um par de roteadores de borda. O provisionamento poderia ser considerado fixo durante todo o caminho entre os roteadores. Tais soluções são interessantes quando uma empresa pretende, por exemplo, emular uma linha virtual entre duas de suas filiais. Soluções desse tipo não podem ser aplicadas a Redes de Núcleo de sistemas celulares, uma vez que os terminais transmissores se movem transmitindo através de diferentes roteadores da borda do domínio.

Atualmente não existem mecanismos de sinalização voltados para um provisionamento dinâmico de recursos em pequena escala de tempo previstos para a arquitetura DiffServ. Além disso, a solução baseada na utilização de provisionamento em excesso é altamente indesejada e deve ser evitada. Assim a utilização da arquitetura DiffServ, principalmente em ambientes de mobilidade, pode levar a violações temporárias dos níveis de QoS contratados [39, 50, 61].

Em ambientes de mobilidade, o controle de tráfego das bordas pode se tornar inadequado com freqüência. Esquemas de alocação dinâmica evitariam tal problema. Tais esquemas necessitam, entretanto, de sinalização por nó e/ou controle de admissão explícito, o que levaria entretanto a uma sobrecarga de sinalização. A partir da sinalização, informações sobre o interior da nuvem DiffServ podem ser levadas às bordas, onde ações de controle de tráfego podem ser tomadas. De fato, a ausência dessa comunicação do estado do interior do domínio DiffServ com suas bordas (onde as decisões de controle de admissão são tomadas) representa uma grande desvantagem dessa arquitetura.

Uma solução possível para as dificuldades de gerenciamento de QoS para o caso de provisionamento estático de recursos, como no caso do DiffServ seria a utilização de sinalização a partir do receptor (*feedback signaling*) e adaptação ou condicionamento dinâmico do tráfego. Assim os eventos de congestionamento causados por provisionamento insuficiente seriam resolvidos através da cooperação de todos, com redução de tráfego (a partir da colaboração das fontes ou através da redução das taxas de suavização dos agregados de tráfego). Um exemplo desse tipo de mecanismo é encontrado no controle de congestionamento do TCP. Para isso os contratos de QoS, os SLA's devem prever tal possibilidade de redução de banda, por exemplo, nos casos excepcionais de escassez de recursos [62].

De uma forma geral, o DiffServ pode ser visto como uma solução que utiliza um provisionamento de recursos de forma totalmente estática, onde o gerente da rede a dimensiona adequadamente para que as necessidades de QoS das classes planejadas sejam alcançados. Para que se atenda de forma satisfatória a todos os usuários e em todos os instantes, esta opção pode exigir grandes desperdícios de banda passante, ou uma perda de generalidade. Isso ocorre porque os recursos da rede devem ser alocados baseados nos possíveis picos de demanda, para que se obtenham garantias absolutas de atendimento aos requisitos de QoS.

3.2 Gerenciamento dinâmico *vs.* propriedade escalar

As propostas de QoS apresentadas na seção anterior podem ser vistas como paradigmas opostos em termos de gerenciamento de recursos. O IntServ oferece garantias de QoS rígidas para cada micro-fluxo do usuário, utilizando uma abordagem onde os recursos são alocados dinamicamente por chamada. Assim, ou se obtém as reservas requisitadas - e nesse caso quem requisita não precisa mais se preocupar em monitorar o nível de serviço - ou a rede indica que não pode oferecer as garantias requisitadas [63]. Tal solução exige uma elevada carga de sinalização na rede, causando sobrecarga de sinalização [39]. Faz-se necessária uma sinalização sempre que se deseje utilizar a rede. Além disso, é preciso que se mantenha um estado por micro-fluxo. Esses dois fatores acabam causando problemas de propriedade escalar.

O DiffServ, por outro lado, busca solucionar o problema de atendimento aos fluxos em escala agrupando os micro-fluxos das aplicações em poucos agregados de tráfego de acordo com suas características comuns. A rede deve então estar bem provisionada, de acordo com as especificações de cada agregado de tráfego. Tal provisionamento é fixo, ou realizado em escalas de tempo muito maiores do que as escalas necessárias para que se acompanhem as mudanças nos padrões e volume do tráfego, o que pode levar a violações temporárias nos níveis de QoS oferecidos aos usuários. Se a rede não for capaz de honrar o pedido, simplesmente não o faz e não notifica isso a aplicação [63].

Os modelos de provisionamento de QoS do IntServ e do DiffServ podem ser considerados extremos opostos de modelos de controle, em termos de propriedade escalar e acurácia. O fator de mobilidade traz dificuldades para ambas as soluções.

A movimentação dos nós exigiria freqüente renegociação de recursos no caso do IntServ, e também tornaria ainda mais difícil a realização de um provisionamento estático adequado, no caso da utilização do DiffServ.

Uma vez que soluções que exijam sinalização e manutenção de um estado por fluxo ou que se baseiem em provisionamento puramente estático não são satisfatórias para que se obtenha QoS fim-a-fim em redes IP, soluções intermediárias podem ser interessantes para as Redes de Núcleo de sistemas celulares 3G.

Dessa forma, soluções intermediárias vêm sendo estudadas, como a propostas de sinalização e reservas por agregação de tráfego, incluindo a proposta de agregação de reservas RSVP [64] que busca reduzir os problemas de atendimento de fluxos em larga escala do RSVP e a arquitetura "Gerenciamento de Recursos para DiffServ" (*Resource Management in DiffServ* - RMD) [65] que propõe esquemas de sinalização para o DiffServ. Outra proposta que esta sendo amplamente discutida é a utilização em conjunto do DiffServ e do IntServ (IntServ sobre DiffServ) [55].

3.2.1 IntServ sobre DiffServ

A arquitetura de IntServ sobre DiffServ permite maior flexibilidade do que ambas quando aplicadas isoladamente, podendo ser aplicada de duas formas diferentes. Na primeira delas, os recursos na rede DiffServ são alocados de forma estática e os equipamentos do domínio DiffServ não reconhecem protocolos de sinalização como RSVP. Os nós de borda, entretanto, são constituídos de duas porções: uma parte do nó faz interface com a rede IntServ na rede do cliente e a outra porção é a interface para a rede DiffServ. Cada um dos nós de borda mantém uma tabela com as capacidades provisionadas de acordo com os contratos para cada serviço DiffServ. A partir dessa tabela é possível que se realize procedimentos de controle de admissão para os fluxos que cruzarão o domínio DiffServ [61]. Essa abordagem permite uma sinalização entre a rede DiffServ e o usuário das redes de acesso. Os nós de borda do domínio, entretanto, continuam sem ter como detectar os congestionamentos no interior do domínio DiffServ.

A segunda estratégia possível é baseada na alocação dinâmica de recursos no domínio DiffServ. O trabalho apresentado em [55] sugere a utilização de roteadores DiffServ compatíveis com o RSVP. Essa abordagem traz as desvantagens de propriedade escalar encontradas no RSVP, que incluem manutenção de estados por micro-fluxo em cada um dos roteadores intermediários e alta carga de sinalização. Outro problema é o tempo de resposta necessário para a negociação das chamadas que pode ser muito longo para determinados requisitos de QoS.

3.2.2 Agregação de reservas RSVP

A utilização de agregações de reservas RSVP foi proposta para melhorar o comportamento do atendimento de largas escalas de fluxos. Assim, nos enlaces de Redes de Núcleo, diversas reservas individuais de RSVP que possuem características comuns de QoS e que dividam o mesmo par de roteadores de entrada e saída são "encapsuladas" na reserva de um agregado. Dessa forma, os nós do interior do domínio trata um número reduzido de agregados de fluxos, o que permite uma diminuição drástica no número de estados armazenados e na sobrecarga de sinalização. Os nós de borda são responsáveis por mapear os pedidos de reservas em um dos agregados existentes ou por criar novas agregações, caso seja necessário.

Uma vez que os fluxos são tratados em conjunto e que não é possível distinguí-los, um mesmo nível de QoS precisa ser garantido a todos de forma que se respeite os requisitos do tráfego mais exigente. Esse fato pode fazer com que as reservas realizadas estejam bastante acima do necessário, representando um desperdício de recursos. A utilização de um maior número de agregações diminui esse problema, pois permite uma maior granulosidade na classificação dos micro-fluxos, aumentando entretanto a sobrecarga de sinalização e a quantidade de recursos armazenados nos elementos de rede. Assim existe um compromisso entre a realização de super-provisionamento e o atendimento a largas escalas de fluxos [66].

O tamanho da reserva de cada agregado pode ser realizado de diferentes formas. A maneira mais simples seria a realização de reservas iguais a soma dos pedidos individuais. As reservas poderiam também ser efetuadas baseadas em mecanismos de predição de carga. Alguma histerese pode ser empregada para diminuir a freqüência das renegociações.

3.2.3 Gerenciamento de recursos para DiffServ (*Resource Management in DiffServ* - RMD)

O controle de tráfego realizado pela arquitetura "Gerenciamento de Recursos para DiffServ" (*Resource Management in DiffServ* - RMD) pode ser visto como uma solução que busca melhorar as garantias oferecidas pelo DiffServ, através da utilização de reservas de recursos por agregados, sem comprometer o atendimento de fluxos em grandes escalas. Essa é uma solução voltada para redes de acesso de sistemas móveis, ou seja, é uma solução que precisa lidar com um grande número de fluxos que possuem requisitos rígidos quanto ao atraso. Por ser uma solução otimizada para redes de acesso, algumas simplificações, como a ausência de suporte a tráfego multicast, são levadas em consideração. Dessa forma, a simplicidade do RMD, quando comparada à soluções como a proposta de agregação de reservas RSVP, é uma característica importante.

O RMD separa o problema de reservas no domínio em reservas em cada nó. Para isso são necessários dois tipos de protocolo, o "Reserva Por Nó" (*Per Hop Reservation* - PHR) e o "Reserva Por Domínio" (*Per Domain Reservation* - PDR). O protocolo PHR pode ser visto como uma extensão ao PHB, incluindo gerenciamento dinâmico de recursos nó-a-nó para cada classe nos elementos da rede [67]. Para tomar as decisões de admissão em cada nó, o PHR permite a realização de reservas de recursos por agregado ou a realização de medidas de estimativas da carga.

No caso do PHR baseado em reservas, os recursos são explicitamente alocados por PHB, e o estado flexível da reserva deve ser renovado periodicamente. As reservas mantêm apenas um estado por PHB: o próprio valor da reserva, em termos de unidades de recursos. A decisão de admissão é baseada em um limiar representando a quantidade de recursos disponíveis. Atualmente existe um protocolo PHR definido, o RODA [68] que é baseado em reservas.

O emprego de PHRs baseados em medidas possui a vantagem de permitir uma melhor utilização do meio de transmissão. Nesse caso, o protocolo sinaliza a existência de recursos na rede sem a necessidade de manter reservas nos nós. Para isso, cada nó deve realizar medições de tráfego para cada PHB e a partir dos valores obtidos decidir quanto à admissão de novos fluxos.

O protocolo PDR, por sua vez, funciona como uma extensão aos PDBs definidos na arquitetura DiffServ. Através do PDR é possível realizar um gerenciamento de recursos em nível de um domínio DiffServ. Esse protocolo trata do controle de admissão baseado nas respostas obtidas pelo PHR. O PDR é parte apenas dos nós de fronteira do domínio DiffServ. Além do controle de admissão, o PDR é responsável pelo mapeamento do pedido de QoS externo para um PHB adequado, realizando a comunicação entre o protocolo de reservas externo e o PHR.

De uma forma geral a arquitetura RMD realiza um controle de admissão distribuído, utilizando sinalização nó-a-nó. Ela busca seguir a filosofia do DiffServ, oferecendo um gerenciamento dinâmico para as diferentes agregações de tráfego. A sinalização permite que o conhecimento do estado no interior do domínio DiffServ seja levado às bordas, onde as decisões de controle de admissão podem ser tomadas.

As soluções baseadas em reservas por agregações de fluxos, como as sugeridas na agregação de reservas RSVP e no RMD não resolvem completamente a sobrecarga de sinalização. Os *handovers* entre RNCs, que serão freqüentes em redes 3G (seção 2.3.1), irão fazer com que conexões troquem de roteador de ingresso na Rede de Núcleo (SGSN) constantemente. Essa freqüente flutuação no volume de tráfego implicará em muitas renegociações para o agregado, sendo que cada uma delas têm os mesmos requisitos de desempenho de um esquema de reservas por fluxo, ou seja, não possui propriedade escalar [61]. Soluções para a diminuição no número de renegociações incluem a utilização de histerese, ou seja, limiares de renegociação, para o aumento e para a diminuição de reservas, diferentes e suficientemente afastados de forma a evitar oscilações no valor da reserva quando a rede estiver em condições estáveis. Essa solução, entretanto, não evita que casos temporários de subutilização ou de alocação insuficiente de recursos ocorram.

3.3 Gerenciamento de recursos

O gerenciamento de recursos, fundamental para que se obtenham os níveis adequados de QoS para os usuários, pode ser visto como o conjunto de várias operações em uma rede. Ele tem como objetivo satisfazer os diferentes serviços a serem oferecidos de forma eficiente. Mecanismos que permitam o gerenciamento adequado dos recursos disponíveis incluem controle de admissão, escalonamento, gerenciamento de filas nos roteadores e controle de fluxo [69]. Nas redes 3G, a grande parcela de tráfego com requisitos de tempo real tende a aumentar os desafios do gerenciamento de recursos.

De uma forma geral, a QoS oferecida pela rede pode ser considerada como uma função dos recursos disponíveis e da carga a qual a rede é imposta [63]. Dessa forma, para que se ofereçam classes de serviços com diferentes níveis de QoS, é importante que se possa controlar a carga admitida por classe de serviço, de forma que os recursos alocados para cada classe sejam suficientes no atendimento de seus requisitos. O controle do nível da carga, que é parte dos mecanismos de gerenciamento de recursos, pode ser considerado um dos fatores mais preocupantes de uma arquitetura de QoS.

O controle de admissão, que é um dos elementos mais importantes de controle dos níveis de carga impostos à rede, pode ser realizado de diversas formas. Todas as arquiteturas de QoS apresentadas anteriormente realizam algum tipo de controle de admissão/tráfego. No caso do IntServ, esse controle é realizado nó-a-nó de forma explícita e dinâmica. O DiffServ por sua vez realiza um controle de admissão implícito através de policiamento/condicionamento da agregação de tráfego, respeitando-se os serviços estaticamente contratados e descritos pelos SLAs [61, 63]. Os controles de admissão do DiffServ e IntServ podem ser vistos mecanismos que operam em escalas de tempo diferentes. De acordo com o trabalho apresentado em [70], o gerenciamento de recursos pode ser classificado conforme a escala de tempo de suas operações. Assim, o mecanismo de policiamento de pacotes do DiffServ opera em uma escala de tempo da ordem do tempo de transmissão de um pacote. Já os mecanismos de controle de admissão de chamadas empregados no IntServ podem ser classificados em uma escala de tempo maior, de chegada de chamadas.

Para que se realize um controle de admissão explícito no DiffServ, assegurando-se QoS a um dado serviço, deve se garantir que a cada salto na rede os recursos alocados para o PHB em questão sejam suficientes para lidar com o total de fluxos. Há então uma necessidade de se estimar a utilização de cada enlace da rede para que se possa otimizar o gerenciamento dos recursos. Mecanismos baseados em medidas realizadas no agregado podem ser utilizadas de forma a otimizar o controle de admissão. Na próxima seção algumas possibilidades de controle de admissão que sejam de alguma forma compatíveis com a filosofia DiffServ (mantendo a propriedade escalar) serão abordadas.

3.3.1 Policiamento de tráfego como controle de admissão

O policiamento de tráfego nas bordas do domínio pode ser considerada como a solução de controle de admissão/tráfego utilizada pela arquitetura DiffServ. Esse controle de admissão funciona de forma relativamente simples, garantindo que o provisionamento estático dentro da nuvem DiffServ se mantenha adequado para cada classe de serviço. Isso é alcançado através da limitação do volume de tráfego admitido nas bordas para cada classe.

O controle de admissão através do policiamento de tráfego nas bordas é utilizado no DiffServ por ser uma solução simples e que tem a capacidade de lidar com grande quantidade de fluxos, apresentando propriedade escalar. As principais desvantagens desse tipo de solução incluem a necessidade de realização de um provisionamento alocando recursos em excesso no interior do domínio, o que é uma conseqüência da ausência de sinalização entre elementos da rede, e a ausência de sinalização/comunicação entre a rede e as aplicações.

O provisionamento em excesso com desperdícios de recursos (ou super-provisionamento) torna-se necessário já que o domínio em questão deve estar planejado para trabalhar em uma situação de pior caso, onde todos os roteadores de borda estejam admitindo o máximo de tráfego, caso se deseje um serviço sem violações. A ausência de sinalização com as aplicações que geram os fluxos faz com que individualmente os fluxos possam estar sofrendo degradação pela ação dos policiadores, mesmo ao se garantir que dentro da nuvem DiffServ os agregados de tráfego estarão sendo servidos de acordo com o esperado para cada classe. Caso o tráfego gerado pelos micro-fluxos estejam excedendo a taxa contratada, pacotes de dados da mesma classe de serviço serão descartados indiscriminadamente pelo policiador, causando uma degradação global percebida pelos usuários finais. Assim, podem ocorrer situações onde todos os micro-fluxos acabam sendo prejudicados.

Alguns trabalhos, como o apresentado em [71], propõem a reconfiguração dinâmica dos elementos de policiamento nas redes DiffServ. Essas reconfigurações são acompanhadas por modificações no provisionamento no interior do domínio [72]. Essa solução, entretanto, não busca otimizar o número de usuários atendidos com base nas características das aplicações, através da flexibilização dos serviços oferecidos.

O policiamento nas bordas representa então um controle de admissão sem sinalização explícita nos casos em que o tráfego é rejeitado [63]. Características do DiffServ, como a distribuição da degradação da QoS entre as conexões que compõem um agregado, que primeiramente poderiam ser consideradas desvantagens, podem ser exploradas positivamente para otimizar a utilização do canal. Isso pode ser realizado aproveitando-se as características de flexibilidade das aplicações de cada classe, permitindo uma degradação suave do serviço. Nesse caso, é de fundamental importância que se possa controlar o nível de degradação máximo a ser atingido. A degradação suave de QoS é um fator importante que deve ser levado em conta nos ambientes de mobilidade, conforme discutido na seção 2.4.

3.3.2 O estado do interior da nuvem DiffServ

As principais dificuldades encontradas no DiffServ, como a garantia de um provisionamento adequado de recursos em todo o domínio, são decorrentes da ausência de comunicação dos estados do interior do domínio DiffServ para suas bordas, onde o controle de admissão é realizado. Dois pontos são importantes quando se pensa em utilizar a rede DiffServ como parte da arquitetura de QoS fim-a-fim com controle de admissão [73]: (i) o método a ser utilizado pelo domínio/nuvem DiffServ para determinar se recursos suficientes estão disponíveis no seu interior e (ii) o método utilizado para que a rede externa/aplicações dos usuários possam saber, junto a nuvem DiffServ, sobre essa disponibilidade.

Para que o estado do interior do domínio DiffServ chegue até as bordas podem ser utilizados métodos distribuídos, baseados em sinalização fim-a-fim como proposto em [65, 74] ou centralizados, baseado em entidades que possuem o conhecimento do estado nos enlaces do interior do domínio [32, 75, 76].

A utilização de um agente central, apesar das desvantagens típicas de abordagens centralizadas, representa uma solução mais simples do que as distribuídas e tem a vantagem de poder tomar decisões mais precisas, por possuir uma visão global do sistema.

O agente central obtém informações dos roteadores através de protocolos de sinalização e, da mesma forma, se comunica com entidades que queiram negociar o acesso à rede DiffServ. Possuindo informações de todo o domínio ele pode realizar procedimentos de AAA (*Authentication, Authorization and Acounting*) e de gerenciamento de QoS. Uma desvantagem dessa proposta consiste na complexidade adicionada pela existência do agente centralizador e da sinalização explícita, que gera tráfego para fins diferentes da transmissão de dados (*overhead*).

A fim de se melhorar a capacidade de lidar com grande número de fluxos, o estado do interior da nuvem DiffServ não deve ser baseado em informações de cada micro-fluxo. Seguindo a filosofia do DiffServ, o controle de tráfego deve se basear em informações dos agregados de tráfego. Uma opção interessante seria realizar o controle de admissão baseando-se em medidas realizadas nas agregações de tráfego, ou seja, implementando-se um "controle de admissão baseado em medidas". Diferentes técnicas de controle de admissão serão apresentadas na seção 4.3.

Esse capítulo apresentou algumas das principais soluções de QoS, realizando uma análise de suas principais características, buscando avaliar seus pontos fortes e suas deficiências sob o ponto de vista de QoS para Redes de Núcleo de sistemas móveis. O próximo capítulo irá propor soluções de provisionamento de QoS que sejam satisfatórias no âmbito da mobilidade, levando-se em conta os requisitos e as características discutidas.

Capítulo 4

Implementação de QoS em Redes de Núcleo

NO presente capítulo são propostas soluções para prover QoS nas Redes de Núcleo de sistemas 3G. Além de serem capazes de lidar com a mobilidade, essas soluções devem permitir o tratamento de fluxos em larga escala. A seção 4.1 traz uma proposta de criação de serviços flexíveis voltados para ambientes móveis, respeitando-se as características de cada aplicação. Possíveis implementações para um desses serviços são abordadas na seção 4.2. Em seguida, na seção 4.3, são apresentados mecanismos de controle de admissão por chamada que podem ser utilizados em Redes de Núcleo para garantir a QoS desejada. O mecanismo de controle de admissão "Soma Medida", que se adequa ao contexto de QoS em mobilidade, é detalhado na seção 4.4. Por fim, a utilização do controle de admissão em conjunto com um dos serviços propostos é apresentado na seção 4.5.

4.1 Serviços flexíveis

Uma forma de se obter QoS voltada para ambientes de mobilidade seguindo a filosofia do DiffServ, evitando-se a sinalização e a utilização de agentes centrais, é tirando-se proveito da capacidade de adaptação de alguns tipos de serviços. Assim novas classes de serviços, com requisitos de QoS mais flexíveis (de acordo com as possibilidades de cada tipo de aplicação) podem ser utilizadas. Tal flexibilidade permite também a diminuição das necessidades de super-provisionamento para que se ofereçam serviços com a qualidade exigida pelas aplicações.

Dessa forma, em casos críticos, onde a capacidade da infra-estrutura é excedida, diferentes tipos de degradação de serviços podem ser oferecidos aos nós móveis. Em casos onde a degradação não fosse permitida algum método de controle de admissão se faria necessário. Esse controle de admissão tem o papel de negar o acesso de novos fluxos quando a demanda ultrapassar níveis determinados, o que tornaria o provisionamento inadequado. Em situações menos rígidas os recursos poderiam ser degradados até níveis mínimos. Alguns serviços poderiam sofrer reduções na banda passante disponível. Opcionalmente, outro tipo de serviço poderia sofrer degradação através de maiores taxas de perdas, ou até mesmo maiores atrasos e *jitter*. O fato do terminal estar parado ou se movendo também pode ser um fator para a degradação do serviço prestado [46], uma vez que garantir QoS após um *handover* é uma tarefa difícil, que envolve reservas de recursos antecipadamente. Essas reservas são recursos que ficarão na grande parte do tempo inutilizados, não sendo interessantes para o provedor de serviço.

A maioria das aplicações possuem requisitos elásticos em diferentes parâmetros de QoS. Para as aplicações interativas em tempo real, como telefonia por exemplo, os baixos atrasos na transmissão dos pacotes apresentam maior impacto na QoS percebida pelo usuário final do que taxas de perdas nulas. Nesse caso, é possível realizar uma solução de compromisso, trocando taxa de perdas por atraso.

É possível então que se combinem a utilização dos elementos previstos na arquitetura DiffServ para que se obtenham as classes que respeitem as características de cada aplicação. Policiadores de tráfego, condicionadores e escalonadores podem ser configurados de modo a otimizar os requisitos de cada classe e ao mesmo tempo maximizar a utilização da rede.

4.1.1 Classes de serviço

Conforme discutido na seção 2.4, a degradação suave do serviço deve ser levada em consideração como uma opção de QoS. Neste contexto aparecem como parâmetros de QoS: a probabilidade de conexão contínua e o perfil de perdas [6, 7]. A partir dos parâmetros de QoS clássicos (atraso, variação no atraso - *delay jitter*, vazão e taxa de perdas) e acrescentando-se a probabilidade de desconexão como fator de QoS, além de se considerar as características de elasticidade das aplicações, foram especificadas classes de serviço voltadas para mobilidade.

Assim, classes de serviço que permitissem certos graus de flexibilidade foram criadas, visando-se especialmente os ambientes de comunicações móveis. Os autores de [32] sugerem uma classificação baseada em apenas dois parâmetros: a taxa de perdas e o atraso. Ao se considerar os quatro parâmetros clássicos de QoS e ainda o fator de conectividade, podem ser derivadas novas classes, atendendo não só as características das aplicações mas também as novas necessidades impostas pelo ambiente de mobilidade. A tabela 4.1 traz essas classes, conforme foi apresentado no trabalho publicado em [77].

A classe para tráfego de controle em tempo real e de informações de emergência em geral (classe A) é uma classe crítica onde não são admitidas perdas ou atrasos significativos. A banda passante necessária, além de ser uma pequena fração do total disponível, é em geral bem conhecida, uma vez que tal tráfego é previsível [32]. A urgência desse tipo de tráfego exige alta conectividade, levando as probabilidades de desconexão a serem nulas, no que depender de provisionamento. A classe B, para aplicações interativas e de tempo real como áudio, vídeo, telefonia e vídeo conferências, tem como requisito a baixa latência. Essas aplicações são, em geral, robustas a determinados níveis de taxa de erros. A banda passante utilizada pode ser alta, em casos de vídeo por exemplo, ou média como para telefonia. O *jitter* deve ser mantido pequeno para que se evitem *buffers* de recepção muito grandes. A probabilidade de desconexão deve ser mantida baixa, evitando-se interrupções no serviço. Ao se pensar em aplicações como login remoto (classe C) por exemplo,

Tabela 4.1: Classes de serviço.

Classe	Aplicação	Taxa de perdas	Atraso	Banda passante	*Jitter*	Probabilidade de desconexão
A	controle em tempo real	nula	baixo	–	–	nula
B	multimídia em tempo real	baixa-média	baixo	média-alta	baixo	baixa
C	login remoto	baixa	baixo	baixa	–	nula-baixa
D	FTP	baixa-média	alto	alta	–	média-alta
E	*web*	baixa-média	médio	média	–	baixa
F	*e-mail, news*	baixa	alto	baixa	–	média-alta

obtém-se uma classe onde atrasos moderados são aceitáveis, assim como taxas de perda medianas. Tais aplicações têm como característica a utilização de pouca banda passante. A conectividade do serviço deve ser alta para o conforto do usuário. Já as aplicações de transferência de arquivos (classe D) apresentam características de altas vazões, com menos importância para atraso. Se existirem esquemas de recuperação de transmissões interrompidas, a conectividade deixa de ser um fator primordial e as probabilidades de desconexão podem ser mais elevadas. Uma classe E, adequada para aplicativos de navegação na *Web*, por sua vez, exige banda passante moderada, mantendo as necessidades de interatividade. Aplicações de pequena vazão e que não exigem interatividade, como *e-mail* e *news*, podem pertencer a uma outra classe de serviço, (classe F), que exige confiabilidade. Mais uma vez, a probabilidade de desconexão alta não influenciará na satisfação do usuário.

O serviço da classe A, por exemplo, é equivalente ao serviço de “linha privada virtual” oferecido pelo pelo PHB EF. Esse serviço teria que ter tráfego pequeno, onde o seu máximo pudesse ser previsível. Desse modo seria possível garantir o correto provisionamento de recursos, incluindo reservas adequadas para que mesmo com a realização de *handovers* parâmetros de QoS, como conectividade, continuem

Tabela 4.2: Classe B, de tempo real em mobilidade.

Classe	Situação	Taxa de perdas	Atraso	Banda passante	*Jitter*	Probabilidade de desconexão
B0	fontes transmitindo antes de sobrecarga	nula	baixo	média-alta	baixo	baixa
B1	fontes gerando sobrecarga	baixa-média	baixo	média-alta	baixo	baixa

garantidos. Isso implicaria em um alto nível de super-provisionamento. Como tal tráfego deve ser uma pequena porcentagem do total e com utilização restrita, isso não representará grandes perdas.

A classe B oferece características de tempo real, com baixo atraso. Por causa de características da mobilidade, é necessário que exista um serviço mais flexível. Dessa forma essa classe será dividida em duas sub-classes, conforme apresentado na tabela 4.2, ambas mantendo as características de baixos atrasos, porém uma delas apresentando maior flexibilidade em termos de taxa de perdas. Com isso, um maior número de conexões pode ser aceita, sendo que o único parâmetro afetado seria a taxa de perdas, ficando o baixo atraso garantido. Essa classe será estudada detalhadamente na seção 4.1.2.

O tráfego para classe C é parecido com o da classe B, exceto pela menor taxa de perdas e menor banda passante. Uma opção seria tratá-la com uma "classe EF de menor prioridade" com uma alocação de recursos mais relaxada, não sendo realizada pela taxa de pico de transmissão. Ocasionalmente iriam ocorrer aumentos de atraso, em caso de rajadas.

Para prover as características da classe D, os serviços oferecidos pelo AF seriam adequados. Nessa classe, a conectividade não é um fator importante. Dessa forma, para que se mantivessem os níveis de serviço, conexões poderiam ser aceitas somente se houvesse taxas mínimas garantidas.

Aplicativos de navegação na *Web*, pertencentes a classe E, possuem característica de tráfego de curta duração. Assim o PHB AF, devidamente sintonizado para tráfegos curtos, seria adequado. A classe F poderia ser atendida por mecanismos de melhor esforço, por não ser exigente em relação aos parâmetros de QoS.

4.1.2 Classe B: tempo real em mobilidade

No serviço da classe B, que exige tempo real porém onde o tráfego não é tão previsível como na classe A, a mobilidade do nó é um fator que influencia a qualidade de serviço oferecida. O mais importante nessa classe é que a característica de baixos atrasos deve sempre ser mantida, em conjunto com a baixa probabilidade de desconexão nos casos de sobrecarga. A qualidade em termos de taxa de perdas, entretanto, pode ser flexibilizada. Assim, caso o nó esteja parado em uma célula, seu serviço de tráfego sem perdas e com baixo atraso estaria garantido. Ao realizar *handover*, entretanto, o nó móvel enfrentaria uma probabilidade de mudar de serviço, para uma sub-classe sem garantias quanto a taxas de perdas. Essa probabilidade iria depender da quantidade de recursos reservados para a classe B. Isso é uma forma de se evitar um alto super-provisionamento com grande desperdício de recursos.

Desta forma, propõe-se que clientes do serviço B que estejam mudando da área servida por um 3G-SGSN para uma adjacente, chegando a uma região com escassez de recursos, utilizem uma sub-classe B1 sujeita a maiores taxas de perdas. Essa sub-classe seria mais flexível permitindo que um nó, mesmo estando parado, sofra degradação até determinados níveis em sua qualidade. Essa degradação ocorreria por causa do aumento do número de nós dentro de uma dada região. Como o número total de clientes atendidos deve ser maximizado para que a probabilidade de interrupção da chamada seja baixa, o número total de clientes da classe B não pode ser facilmente estabelecido e sobrecargas podem ocorrer. Durante esses períodos de sobrecarga, clientes da classe B que já estivessem na célula antes dos recursos se tornarem escassos ficariam pertencendo a uma sub-classe prioritária B0, mantendo as características da classe B incluindo baixas perdas. Uma idéia similar é apresentada em [46].

Diferentes opções podem ser aplicadas para implementação dessa classe de tempo real com perdas. Algumas delas serão apresentadas na próxima seção.

4.2 Implementação das classes de serviço

As classes de serviço propostas na seção anterior podem ser implementadas de diferentes maneiras, através da utilização dos mecanismos de controle de tráfego abordados. Escalonadores, policiadores, condicionadores e mecanismos de gerenciamento de *buffer* podem ser combinados para que se obtenham as características de QoS desejadas.

Nessa seção, serão apresentados três mecanismos diferentes (dois escalonadores de pacotes e um gerenciamento de *buffer*) que podem ser utilizados para fornecer os serviços de tempo real das sub-classes B0 e B1, como especificadas na seção 4.1.2. Ambas as sub-classes devem fornecer um serviço de baixo atraso, sendo que a sub-classe B1 apresenta maior tolerância a perdas.

Os mecanismos escolhidos para serem avaliados são os escalonadores: PQ, pela sua simplicidade, o que facilita a interpretação dos resultados; e WRR, que permite a criação de um serviço que respeite estritamente as especificações do PHB EF do DiffServ; além do gerenciamento de *buffer* SPP, por suas características de atendimento de uma classe de maior prioridade otimizando o serviço da classe inferior. Os mecanismos serão detalhados a seguir.

4.2.1 Fila Prioritária - PQ

A utilização de escalonadores de pacotes para implementar as classes de serviço implica na utilização de duas filas FIFO, uma para cada classe, conforme o esquema apresentado na figura 4.1. O algoritmo de escalonamento irá decidir de qual fila o pacote será enviado.

O algoritmo de Fila Prioritária (*Priority Queue* - PQ) é o mais simples que pode

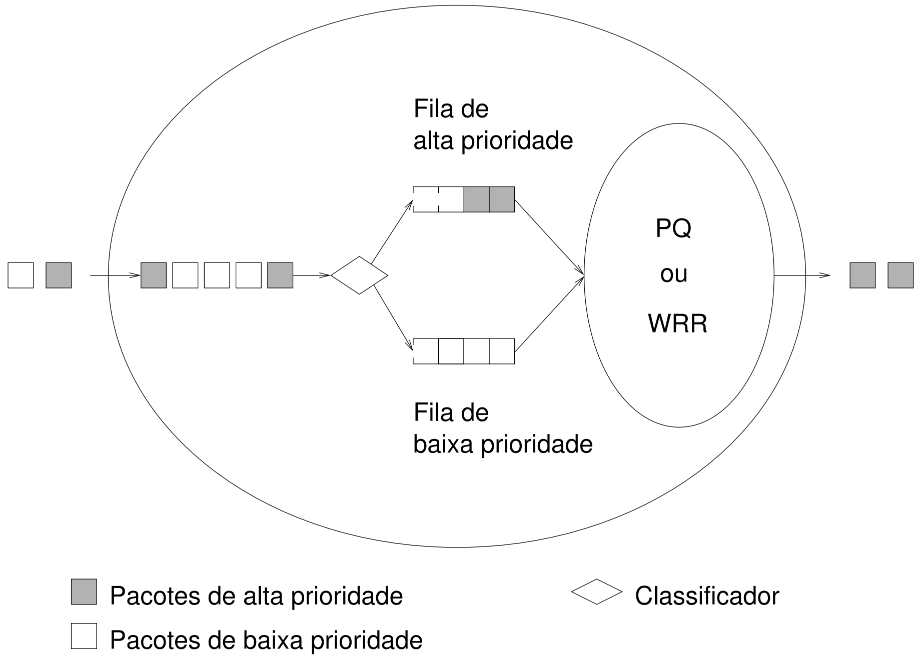


Figura 4.1: Nó que implementa diferenciação de classes através de utilização de múltiplas filas e escalonadores de pacotes.

ser utilizado pelo escalonador de pacotes na implementação o serviço em questão. Ele é baseado em múltiplas filas FIFO que são escalonadas de acordo com a prioridade do tráfego.

Os agregados de tráfegos das sub-classes B0 e B1 são marcados nas bordas da Rede de Núcleo. Cada uma das agregações será então classificada para a fila correspondente no roteador. Ao transmitir, um escalonador de prioridade (*Priority Queue* - PQ) escolhe de qual fila será enviado o próximo pacote. Assim, sempre que houver pacote de tráfego B0, este será enviado, e os pacotes de tráfego B1 somente poderão ser transmitidos caso a fila prioritária esteja desocupada. Este esquema possui como principal vantagem a simplicidade de implementação, podendo entretanto causar grandes atrasos ou até mesmo *starvation* na transmissão de pacotes menos prioritários[1]. Essa solução, portanto, garante o desempenho do tráfego prioritário, sem oferecer garantias ao tráfego de menor prioridade.

Policiadores de tráfego podem ser utilizados em conjunto para limitar o volume

[1] O fluxo menos prioritário nunca teria seus pacotes enviados.

da sub-classe B1. Essa opção, ao limitar a carga, permite que os atrasos observados se mantenham controláveis.

4.2.2 *Round Robin* com peso - WRR

Outro escalonador a ser avaliado dando suporte às classes propostas é baseado no algoritmo *Round Robin* com peso (*Weighted Round Robin* - WRR) [9], que é amplamente conhecido e comumente empregado em arquiteturas de diferenciação de serviços. Filas FIFO duplas são utilizadas, conforme apresentado na figura 4.1, de forma similar ao esquema com escalonador de prioridade. O escalonador WRR aloca uma fatia de serviço no enlace de saída para cada fila. Cada classe de serviço deve então ser dimensionada estaticamente de acordo com a porção de banda passante alocada. Ocorre então uma reserva de banda passante por classe. O serviço da sub-classe B0, por exemplo, deve ser baseado no PHB EF. Para que a definição do EF sempre seja válida será necessário que se utilize e configure corretamente os suavizadores/policiadores de tráfego para o agregado. Assim, os suavizadores devem garantir que as taxa de chegada de pacotes em cada fila seja menor que suas respectivas taxas de serviço, alocada pelo WRR. No caso de sistemas 3G UMTS, a suavização será realizada no SGSN e no GGSN, que são os roteadores de ingresso para a Rede de Núcleo baseada em DiffServ.

A configuração dos suavizadores, indicando taxa máxima e rajada, representa uma solução de compromisso entre o nível de QoS obtido para os diferentes parâmetros de interesse e a porcentagem de reserva de banda necessária para que as classes sejam atendidas de forma satisfatória. Uma configuração estática que atenda aos requisitos de QoS é possível em ambientes fixos, ficando dificultada nos ambientes de mobilidade, que apresentam variações no volume de tráfego transmitido. No caso da classe B, o provisionamento adequado seria sempre garantido para a porção dos clientes na sub-classe B0. O número total de clientes da sub-classe B0, entretanto, não pode ser facilmente estabelecido. Assim, os clientes B1 que estiverem gerando uma situação de sobrecarga estariam suscetíveis a degradação na taxa de perdas. A configuração do conjunto suavizador/escalonador para a sub-classe B1 deve então garantir que características de interesse (no caso, baixos atrasos) sejam respeitadas, mesmo em situações de sobrecarga, quando outros requisitos de QoS podem sofrer degradação (como a taxa de perdas). Essa configuração envolve escolha de tamanhos de *buffers*, de taxas de suavização, e porcentagem do enlace reservada a cada classe, sendo uma tarefa complexa.

Mesmo nas situações de sobrecarga, o serviço da sub-classe B1 ainda seria baseado no PHB EF. O tráfego em excesso seria descartado/suavizado nas bordas do domínio. Assim, em termos dos elementos da arquitetura DiffServ, o serviço oferecido pela sub-classe B1 seria baseado no PHB EF, possuindo entretanto um comportamento por domínio, o PDB, que permite a ocorrência de perdas, descartando o tráfego que excede o provisionamento para o EF. A opção por baixos atrasos em troca de maiores taxas de perdas é configurada através da parametrização dos condicionadores, através da escolha adequada do tamanho dos seus *buffers*.

4.2.3 Gerenciamento de *buffers* SPP

Além da utilização de escalonadores, como o PQ e o WRR, a classe de tempo de real poderia ser implementada a partir de políticas de gerenciamento de *buffer*. Nesse caso, apenas uma fila é compartilhada pelos pacotes de todas as classes, mas o envio e o descarte de pacotes são realizados de forma diferenciada. Assim, uma opção para a implementação de classes de tempo real com perdas concorrendo com tráfego prioritário, sem perdas, é a utilização de políticas protetoras para esse gerenciamento de *buffers*. Tais políticas garantem o serviço da classe prioritária oferecendo porém garantias para a classe inferior. Um exemplo de tais políticas, a "Política Protetora Simulada" (*Simulated Protective Policy* - SPP), é apresentado em [8]. Assim, propõe-se a utilização de tal política no contexto de diferenciação de serviços em ambientes móveis, para o suporte da classe B.

Uma política de serviço de pacotes é dita protetora quando ela consegue manter as garantias do tráfego prioritário, como taxa de perdas, independentemente da carga e dos padrões de chegada do tráfego de menor prioridade. Tais políticas são importantes nas condições em que torna-se difícil a tarefa de prever o comportamento do tráfego menos prioritário, como no caso das redes móveis. O mecanismo de SPP possui ainda uma característica que a faz importante para o cenário estudado, onde o desempenho do tráfego B1 não deve ser deixado de lado. O SPP é uma política que garante o desempenho mínimo para o tráfego mais prioritário (B0, no caso de redes móveis aqui apresentado) mas também otimiza o desempenho do tráfego menos prioritário B1.

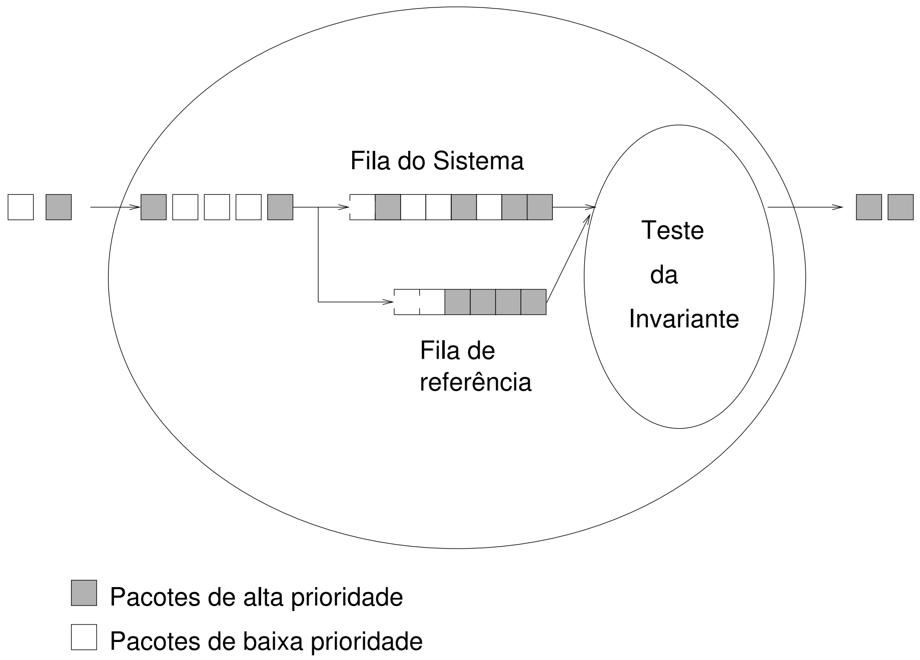


Figura 4.2: Nó com o mecanismo SPP.

O SPP funciona simulando uma fila de referência que representa uma fração do *buffer* do nó. Os pacotes mais prioritários que chegam são armazenados no *buffer* e também na fila simulada. Os pacotes menos prioritários só podem ocupar o *buffer*, não entrando na fila simulada. Caso o *buffer* esteja cheio e um pacote mais prioritário chegue, um menos prioritário deverá ser descartado caso ainda haja espaço na fila simulada. No momento do envio de pacotes uma invariante também precisa ser respeitada: o número de espaços disponíveis no *buffer* para pacotes de alta prioridade deve sempre ser maior ou igual ao espaço disponível na fila de referência. Deve se notar que os espaços ocupados por pacotes de baixa prioridade contam como espaços vazios para os de maior prioridade, uma vez que são descartados em caso de necessidade (através de técnicas de *push-out*). Essa invariante deve ser verificada ao se tentar servir pacotes de menor prioridade. Caso ela seja violada, o pacote de baixa prioridade que seria enviado deve ser descartado.

Propõe-se então que a política SPP seja utilizada para prover os serviços desejados pelas sub-classes B0 e B1, evitando perdas para a sub-classe B0 e buscando otimizar o atendimento da sub-classe B1. A figura 4.2 apresenta este esquema. Os pacotes das diferentes classes são marcados na entrada da Rede de Núcleo. A fila SPP trata diferenciadamente os pacotes.

4.3 Controle de admissão por chamada

Conforme discutido na seção 3.3, o controle de admissão é parte fundamental dos mecanismos de controle de tráfego de gerenciamento de recursos. O controle de admissão é importante para que se garanta que o provisionamento da rede se mantenha adequado. Dessa forma as situações de sobrecarga que poderiam eventualmente ocorrer, causadas por um excesso de usuários acessando a rede, podem ser controladas. Os mecanismos de controle de admissão permitem então que se mantenham os níveis de QoS dos fluxos admitidos dentro das especificações desejadas, evitando a deterioração da qualidade de serviço dos fluxos já estabelecidos ao se admitir um novo fluxo.

Uma forma simples de se garantir compromissos de desempenho é permitir a admissão de fluxos baseando-se em uma situação de pior caso. Um exemplo é o controle de admissão baseado nas taxas de pico dos transmissores. Essa política de controle de admissão pode se mostrar bastante ineficiente em algumas situações, como por exemplo em um cenário que possua fontes de tráfego com características de transmissão em rajadas. Nesse caso, tal estratégia pode resultar em uma baixa utilização da rede.

Por esse motivo, métodos sofisticados de controle de admissão vêm sendo estudados pela comunidade de pesquisa em redes [78]. Tais métodos procuram otimizar a relação entre alta utilização dos recursos e atendimento das necessidades de cada fluxo dos usuários. Para isso são utilizadas modelos analíticos que descrevem o comportamento dos tráfegos e a entidade de controle de admissão utiliza-se de propriedades estatísticas para a tomada de decisão [79].

Existem basicamente duas abordagens para realização do controle de admissão de chamadas. O controle de admissão pode ser baseado em parâmetros ou baseado em medidas [69].

A primeira abordagem, baseada em parâmetros (*Parameter Based Admisson Control* - PBAC) calcula a quantidade de recursos disponíveis na rede com base nos recursos já consumidos pelos fluxos admitidos. A decisão de admissão é tomada de acordo com a disponibilidade atual e com as necessidades do novo fluxo. Para tanto a rede necessita de uma descrição precisa, conhecida *a priori*, do comportamento do tráfego de cada um dos fluxos. É também necessário que a rede mantenha estados por fluxos para que possa realizar cálculos de decisão de admissão de novos usuários.

Já no segundo caso, controle de admissão baseado em medidas (*Measurement Based Admisson Control* - MBAC) [80], a rede toma suas decisões de admissão baseando-se em medidas atuais da carga de tráfego. Um descritor simplificado do novo fluxo, sem necessidade de alta acurácia e precisão, e medições atuais da rede são utilizados na decisão de admissão. Não há necessidade de utilização de descritores de tráfego complexos que modelem de forma detalhada os novos fluxos a serem admitidos, como os utilizados no PBAC. Tal descritor só será utilizado no instante da admissão. Se ele estiver equivocado, não haverá grandes problemas, uma vez que os próximos fluxos a serem admitidos não se basearão no modelo matemático, mas sim nas medições do conjunto de fluxos.

4.3.1 Controle de admissão baseado em medidas - MBAC

Esquemas de controle de admissão baseados em medidas possuem muitas vantagens quando comparados aos esquemas baseados em parâmetros. Primeiramente, a decisão de admissão é realizada com base na utilização real dos recursos. Em oposição ao que ocorre nos casos de controle baseado em parâmetros, o descritor de tráfego dos fluxos que chegam não precisa possuir uma acurácia elevada uma vez que sua real influência no agregado será medida. Assim, as estimativas de tráfego são somente utilizadas para a decisão de admissão do próprio fluxo. Os próximos fluxos que venham a ser admitidos não irão sofrer influência de erros de super-estimativas desses fluxos. Essa característica ajuda na otimização da utilização dos recursos, diminuindo as chances de sub-utilização dos enlaces.

Outra característica importante do MBAC é a reduzida carga de mensagens de sinalização. Não há necessidade de se utilizar sinalização explícita quando uma fonte finaliza uma transmissão. Os recursos estarão automaticamente liberados, uma vez que novas medições realizadas irão refletir a queda na utilização da rede. Nos esquemas baseados em parâmetros ou em reservas, torna-se necessária a notificação explícita do fim da utilização ou o emprego de mensagens periódicas que indiquem a continuação do uso. No MBAC, a única variável de estado guardada nos roteadores será a carga medida que cada classe está impondo à rede. Outro fator interessante no MBAC é que, em oposição ao que ocorre no PBAC, a entidade que realiza as decisões de admissão poderá se basear em medidas realizadas diretamente sobre os agregados de cada classe de serviço, o que está de acordo com a filosofia DiffServ.

No entanto, mesmo que a carga na rede se mantenha estável, o estado futuro dos fluxos admitidos e a carga que os mesmos impõem à rede podem se modificar. Assim, as decisões de admissão realizadas com dados atuais podem não ser adequadas para situações futuras. Essa característica desqualifica o MBAC para garantir a oferta de QoS em termos estritos, sendo o mesmo mais apropriado para que se forneça serviços a aplicações adaptativas, através de garantias estatísticas de QoS. A ausência de garantias rígidas de QoS não é grave em um ambiente móvel, onde as aplicações devem ser capazes de lidar com certos níveis de degradação de serviços. O MBAC, por sua vez, é capaz de perceber as variações causadas no tráfego da rede causadas pela mobilidade dos nós, utilizando-se, para tanto, de uma baixa carga de sinalização. O MBAC é então a melhor opção para a realização de controle de admissão em ambientes de mobilidade [69, 81] que se baseiam na arquitetura DiffServ, como o caso das Redes de Núcleo de sistemas 3G.

Há entretanto a necessidade de haver uma comunicação entre os elementos interiores da rede com aqueles que realizam o controle de admissão, como discutido na seção 3.3.2. A decisão de admissão pode então ser tomada de forma distribuída, como proposto na arquitetura RMD, onde os nós participam de uma sinalização fim-a-fim, ou então de forma centralizada, onde os nós transmitem seu estado para uma entidade central. A maior preocupação da segunda opção é a capacidade da entidade central lidar com um grande número de elementos de rede, ou seja, possuir propriedade escalar. Muitos trabalhos vêm sendo realizados em busca de uma arquitetura que atenda esse requisito. Por exemplo, estruturas hierárquicas baseadas em domínios lógicos foram propostas em [32, 75] e uma arquitetura em duas camadas é estudada em [76].

O trabalho realizado aqui faz uso de MBAC para permitir o controle dos níveis de degradação que venha a ocorrer em situações de sobrecarga. O MBAC se mostra adequado para construir uma estrutura de serviços flexíveis para tráfego de tempo real em ambientes de mobilidade quando utilizado em conjunto com uma arquitetura baseada em DiffServ, empregando escalonadores e condicionadores/policiadores adequados. O MBAC deve permitir uma alta utilização dos recursos e manter os níveis de QoS dentro de padrões desejados.

4.4 Mecanismo de MBAC: Soma Medida

Os mecanismos de controle de admissão baseados em medidas são compostos por dois componentes fundamentais: o processo utilizado para realizar as medidas de estimativa da rede e o critério empregado para decidir sobre a admissão, baseado nas medições realizadas [82].

Através da combinação de diferentes mecanismos de medidas e da utilização de diferentes critérios de decisão, múltiplas combinações podem ser realizadas resultando em uma diversidade de esquemas de controle de admissão. No caso do controle de admissão para a Rede de Núcleo de sistemas 3G, o mecanismo deve possuir propriedade escalar. Dessa forma, alguns mecanismos de MBAC, como os baseados em medições individuais, ficam excluídos. A simplicidade da comunicação entre os fluxos a serem admitidos e os pontos de tomada de decisão deve também ser levada em conta. Assim, os mecanismos que utilizem descritores de tráfego simples devem ser prioritariamente considerados.

O trabalho apresentado em [82] realiza uma avaliação do desempenho de seis mecanismos que possuem as características descritas acima. Os mecanismos que são abordados e detalhados no trabalho são: *Measured Sum, Hoeffding Bounds, Tangent at Peak, Tangent at Origin, Measure CAC* e *Aggregate Traffic Envelopes*. Eles diferem em seus algoritmos de estimativa e de decisão de admissão. O trabalho conclui que os diferentes mecanismos obtém resultados idênticos. Todos os mecanismos estudados atingiram uma relação entre utilização e taxa de perdas equivalente. Todos as opções avaliadas apresentaram dificuldades semelhantes para a configuração e ajuste do desempenho dos mecanismos.

Uma vez que os métodos de MBAC não apresentam diferenças significativas em seus desempenhos optou-se aqui pela utilização de um algoritmo relativamente simples, que seja de fácil configuração e de baixo custo computacional. Assim, foi escolhida para avaliação a combinação do algoritmo de decisão "Soma Medida" (*Measured Sum*) [83] com o mecanismo de estimativa "Janela de Tempo" (*Time Window*) [84].

O mecanismo de "Soma Medida" é baseado em medições de carga do agregado de tráfego. Novas admissões somente são aceitas se a estimativa da utilização no instante do pedido de admissão somada à carga causada pelo novo fluxo for menor do que um nível máximo de utilização do canal. A equação 4.1 apresenta o teste de admissão descrito:

$$\nu + r^{\alpha} < \upsilon\mu \tag{4.1}$$

onde ν é a carga estimada do tráfego existente, r^{α} é o descritor de tráfego com a taxa prevista para o fluxo α, μ é a capacidade do canal e υ é a porcentagem de utilização máxima do enlace.

O método de medida "Janela de Tempo" realiza estimativas da utilização média do enlace a cada intervalo de tempo S. Ao final de uma janela de tempo T, a maior amostra de utilização desse período será a estimativa empregada para as decisões de admissão durante as medições para a nova janela. O valor da estimativa é atualizada imediatamente em situações de aumento de carga: sempre que a carga medida é superior ao valor anteriormente estimado e toda vez que um novo fluxo é admitido. Nesse último caso a estimativa de utilização é artificialmente aumentada. A carga de uma fonte individual α, indicada pelo seu descritor de tráfego, é somada à utilização medida ($\nu' = \nu + r^{\alpha}$). Cada vez que um novo fluxo é admitido e sua carga é somada à carga medida, a contagem de tempo da janela T deve também ser reiniciada de forma que haja uma janela inteira para que o mecanismo possa medir a real influência desse novo tráfego na carga da rede. O mecanismo está apresentado na figura 4.3.

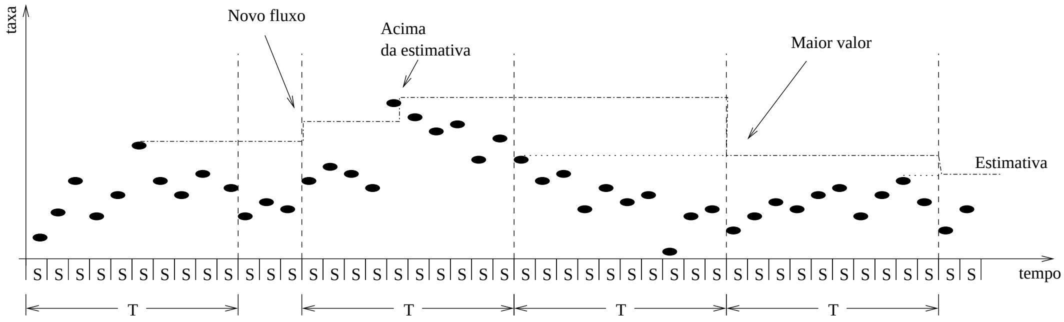


Figura 4.3: Mecanismo de medições "Janela de Tempo".

4.4.1 Sintonia do mecanismo

O desempenho do mecanismo de controle de admissão pode ser medido em termos de taxa de perdas dos usuários e eficiência de utilização do meio. De uma forma geral, pode ser dito que quanto mais conservador é um mecanismo de controle de admissão, menor será a taxa de perdas observada, porém a utilização do meio será baixa. O mecanismo de MBAC apresentado possui alguns parâmetros que permitem a realização de uma sintonia do desempenho desejado [84]. Os parâmetros de ajuste no caso apresentado são: S, T e v.

O período de amostragem S permite o controle da sensibilidade das medidas realizadas. Quanto menor o período, mais sensível a rajadas ficará o estimador. Períodos maiores irão permitir medidas mais suaves. Quando se utilizam valores de S pequenos, as medidas de utilização tendem a refletir a taxa da rajada, tornando o controle de admissão mais conservador.

Ao final de cada janela de tempo T a estimativa de carga é ajustada ao valor medido. Quando a estimativa de carga ν é aumentada, seu valor permanecerá alto pelo menos até o início de uma nova janela de tempo. Assim, o tamanho da janela T controla a capacidade de adaptação do mecanismo para diminuições nos valores estimados. Valores pequenos permitem um melhor aproveitamento do canal, implicando em uma menor estabilidade, podendo ocasionar maiores taxas de perdas. Uma janela de tempo deve permitir a realização de um número significativo de medições de período S. A referência [83] sugere que para redes fixas a relação $T/S \geq 10$ seja mantida.

Quando se deseja implementar um esquema de MBAC é importante que se possa garantir a estabilidade do sistema de filas. Caso a variância do tamanho das filas seja muito grande, os esquemas baseados em MBAC não funcionam corretamente. Essas situações ocorrem quando o sistema aproxima-se de sua utilização máxima. Assim, o fator de utilização máxima do canal, representado pelo parâmetro υ, passa a ter importância. O valor adequado de utilização do enlace depende de características do tráfego. Quando a taxa de cada fonte é pequena quando comparada com a capacidade do enlace e quando as rajadas são curtas, esse fator pode ser escolhido com um valor alto.

De todos os fatores de ajustes apresentados, o que possui maior influência no compromisso utilização *vs.* taxa de erros é o tamanho T da janela de tempo [83].

A influência da escolha dos parâmetros de sintonia do MBAC no seu desempenho para diferentes níveis de mobilidade será avaliada através de simulações na seção 5.4.1

4.5 Controle de admissão para classe B: MBAC em dois níveis

Com o objetivo de limitar o nível de degradação permitido nas classes de serviços flexíveis propostas na seção 4.1.1 é possível que se utilize os mecanismos de controle de admissão. O controle de admissão deve limitar o número de fluxos de cada classe transmitindo na rede, de forma que as especificações de QoS sejam respeitadas.

No caso da classe B (tempo real voltada para mobilidade), por exemplo, propõe-se a utilização de esquemas de controle de admissão em dois níveis. Um primeiro nível, mais conservador, seria responsável pela admissão dos fluxos com tratamento prioritário (B0) em caso de sobrecarga. Na realidade, esse controle de admissão rejeitaria os fluxos que passassem a desrespeitar os limites do provisionamento para tal classe de forma conservadora para garantias mais rígidas de QoS.

Os fluxos que causassem a sobrecarga seriam encaminhados à sub-classe de prioridade mais baixa (B1), conforme descrito na seção 4.1.2. Essa classe de serviço é planejada para lidar com níveis de degradação de QoS em troca de um serviço com menores probabilidades de interrupções. Assim, um controle de admissão menos conservador seria utilizado de forma a maximizar o número de usuários concorrentes, mantendo um nível máximo de degradação.

Diferentes tipos de controle de admissão podem ser empregados em ambos os níveis. Esquemas estáticos, MBAC e PBAC podem atuar, cada qual resultando em desempenhos diferenciados em termos de ocupação do canal e da qualidade dos serviços obtidos. Parâmetros relacionados com a mobilidade dos nós devem influenciar o desempenho de cada esquema.

Ao se utilizar o mecanismo de MBAC "Soma Medida", que se mostra adequado para ambientes móveis (seção 4.4), as configurações para cada classe devem ser realizadas de maneiras diferentes. Assim, para a classe menos prioritária, uma configuração menos conservadora deve ser adotada, permitindo certos níveis de degradação em troca de um maior número de fluxos aceitos, o que resulta em maior conectividade/disponibilidade para os usuários. Essa degradação, entretanto, estaria limitada a níveis conhecidos, permitindo que se ofereça um serviço adequado à ambas as classes.

Nesse capítulo foram apresentadas e propostas implementações de classes de serviço voltadas para Redes de Núcleo de sistemas 3G. O capítulo seguinte apresenta uma avaliação de desempenho de alguns desses esquemas. A partir das simulações realizadas, buscou-se comparar as diferentes possibilidades através de uma análise crítica dos resultados obtidos.

Capítulo 5

Simulações

NESTE capítulo são apresentadas simulações para a análise de desempenho de alguns dos mecanismos discutidos ao longo do texto. Na seção 5.1 há uma breve apresentação do simulador de redes escolhido. O cenário simulado é então descrito na seção 5.2. Em seguida, os mecanismos PQ, SPP e WRR são simulados e analisados, na seção 5.3. Em seguida, na seção 5.4.1, é realizado um estudo para a sintonia do mecanismo de MBAC em ambientes com mobilidade. Por fim, na seção 5.4.2, é avaliada a utilização em conjunto do MBAC com o PQ, SPP e o WRR, para a implementação do serviço da classe B.

5.1 O simulador *ns-2*

Nesse capítulo serão apresentadas as simulações para a análise do desempenho de alguns dos temas abordados ao longo do texto. As simulações são voltadas para a classe de serviço de tempo real em mobilidade (classe B, seção 4.1.2). Elas foram realizadas através do simulador de redes *ns-2* [85]. Esta ferramenta de simulação de redes foi desenvolvida na Universidade da Califórnia em Berkeley, em parceria com o projeto VINT (*Virtual InterNetwork Testbed*), incluindo como colaboradores o USC/ISI, a Xerox PARC e o LBNL. O simulador *ns-2* foi escolhido para as simulações devido a sua ampla aceitação no meio científico e pela disponibilização gratuita de seu código fonte, permitindo que as modificações necessárias fossem desenvolvidas.

O *ns-2* é um simulador de redes orientado a eventos escrito em C++ e que utiliza a linguagem de *script* orientada a objeto OTcl para a interface com o usuário. A utilização de linguagens diferentes permite mais flexibilidade ao simulador. Se o usuário estiver simulando novos protocolos, cujas rotinas complexas gastem muito tempo de processamento, ele pode utilizar-se da eficiência e da robustez do C++ para a manipulação de bytes e de cabeçalhos de pacotes por exemplo. O C++ oferece tempos de execução baixos, porém a depuração pode ser mais lenta.

O OTcl, que é utilizado para a configuração dos cenários de simulação, apresenta características adequadas para esta função. Diferentemente do usuário que busca a implementação de um protocolo complexo, aquele que busca analisar os efeitos de modificações de parâmetros nas redes necessita de facilidade e rapidez na configuração de cenários. Como a configuração de um cenário é realizada apenas no início da simulação, o tempo de execução passa a não ser um fator crítico, sendo a simplicidade da interação com o usuário o aspecto mais importante. Rotinas em OTcl podem ser utilizadas também na simulação de mecanismos que não necessitem lidar com um grande número de operações.

No caso do trabalho aqui apresentado, o mecanismo de escalonamento SPP foi implementado em C++, para que se obtivessem os níveis de desempenho computacional desejados. Para tanto, utilizou-se a descrição do mecanismo efetuada em artigos publicados. Depurações e testes foram realizados no ambiente Linux, utilizado no laboratório do GTA/COPPE. Os mecanismos de MBAC e todo o cenário de mobilidade foram implementados em OTcl por causa da flexibilidade e facilidade de utilização, além da maior agilidade nas reconfigurações necessárias. O mecanismo de MBAC, similarmente ao que ocorreu no SPP, foi implementado e testado baseando-se nas descrições obtidas em artigos publicados. As extensões para dar suporte aos mecanismos de DiffServ que foram utilizadas haviam sido previamente implementadas e validadas por outros membros do grupo de pesquisa do GTA/COPPE [86, 87].

Técnicas de replicações independentes foram utilizadas para que se obtivessem os níveis desejados de intervalo de confiança e de erros. Foram tomados cuidados para que as simulações fossem realizadas com sementes geradoras de números aleatórios diferentes em cada replicação. Os resultados das simulações foram considerados após o descarte de seus transitórios [88].

5.2 Cenário

O cenário de simulações utiliza uma uma topologia que busca representar uma Rede de Núcleo de um sistema celular e está apresentada na figura 5.1. Nesse estudo a preocupação principal não foi modelar o sistema móvel, mas sim avaliar os mecanismos propostos quanto a garantias de atraso e o grau de diferenciação obtido, levando-se em consideração a chegada de novas fontes devido à mobilidade. O atraso foi escolhido como métrica porque ele é um dos fatores mais importantes para QoS de aplicações multimídia em tempo real, as quais são focadas pelas sub-classes B0 e B1 simuladas. A taxa de perdas também é parâmetro de QoS para as sub-classes simuladas, devendo ocorrer proteção da sub-classe B0, ou seja, mesmo que se faça necessária uma degradação em termos de taxa de perdas para a sub-classe de menor prioridade B1, a sub-classe de mais alta prioridade deve permanecer com taxas de perdas próximas a zero.

A modelagem de tráfego de tempo real foi realizada utilizando-se fontes de tráfego *on-off*, com características indicadas na tabela 5.1. Essas fontes apresentam intervalos de atividade exponencialmente distribuídos com média de 1200 ms e intervalos de inatividade com distribuição equivalente, porém com média de 1800 ms. A taxa de transmissão nos intervalos de atividade é de 64 kbps. Essas fontes representam tráfego de voz codificado por PCM com supressão de silêncios [89], sendo transmitidas pela rede IP através do protocolo de transporte UDP. O tráfego gerado por cada fonte apresenta então uma taxa média de 25,6 kbps. Cada pacote de 256 bytes contém 32 ms de voz.

As fontes foram divididas entre de alta prioridade, da sub-classe B0, e de baixa prioridade, da sub-classe B1. As fontes B0 estão conectadas através dos nós de 6

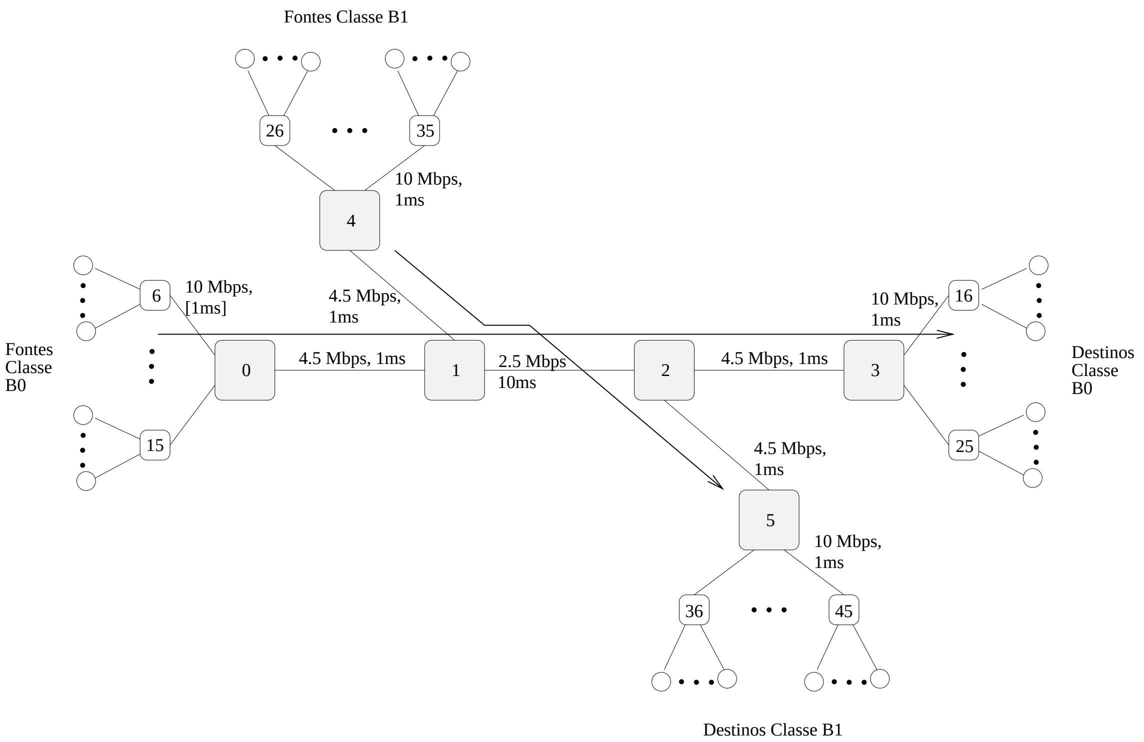


Figura 5.1: Topologia.

a 15, enviando os fluxos aos nós 16 a 25. As fontes B1 se conectam através dos nós 26 a 35 e destinam-se aos nós 36 a 45. O enlace de gargalo, é aquele entre os nós 1 e 2. Os mecanismos descritos na seção anterior atuam junto ao nó 1, na entrada do gargalo. Os nós 0 e 4 poderiam ser vistos como as interfaces de entrada de um único nó, representando o roteador que conecta a rede de acesso sem fio à infra-estrutura fixa (o SGSN na arquitetura UMTS). As características de vazão e atraso dos enlaces estão indicadas na figura.

Tabela 5.1: Modelo de tráfego de voz.

Amostragem	1 amostra de 8 bits a cada 125 μs
Distribuição	*on-off*; on = 1200 ms; off = 1800 ms
Taxa	on = 64000 bps; off = 0 bps
Taxa média	25600 bps
Tamanho do pacote	256 bytes

A capacidade do enlace de gargalo é suficiente para que se possam transmitir simultaneamente 97 fontes, considerando-se a taxa média de transmissão das mesmas. Os enlaces de acesso permitem a transmissão de 175 fontes simultâneas.

5.3 Comparação dos mecanismos WRR, PQ e SPP

Nessa seção é realizada uma comparação dos mecanismos propostos para implementação da classe B, apresentados na seção 4.2. Algumas das simulações aqui apresentadas são semelhantes àquelas realizadas em trabalhos anteriores [77, 90]. Os mecanismos são implementados no nó 1, de entrada do enlace de gargalo. Os demais nós implementam filas FIFO do tipo *droptail*, ou seja, filas que realizam descartes somente quando encontram-se completamente cheias, descartando o último pacote que chegou. Ao se utilizar mecanismos de duas filas no nó 1 (PQ e WRR), ambas possuem o mesmo tamanho, sendo iguais à metade da fila única do SPP. Isso representa um roteador com uma quantidade limitada e fixa de memória, sendo configurado para um dos dois esquemas. Dessa forma, os nós com uma só fila armazenam até 240 pacotes. Nos casos do WRR e do PQ, cada sub-classe possui uma fila de 120 pacotes. O suavizador de tráfego, empregado no esquema baseado em WRR, foi aplicado ao nó 0, para o tráfego B0, e ao nó 4, para o tráfego B1.

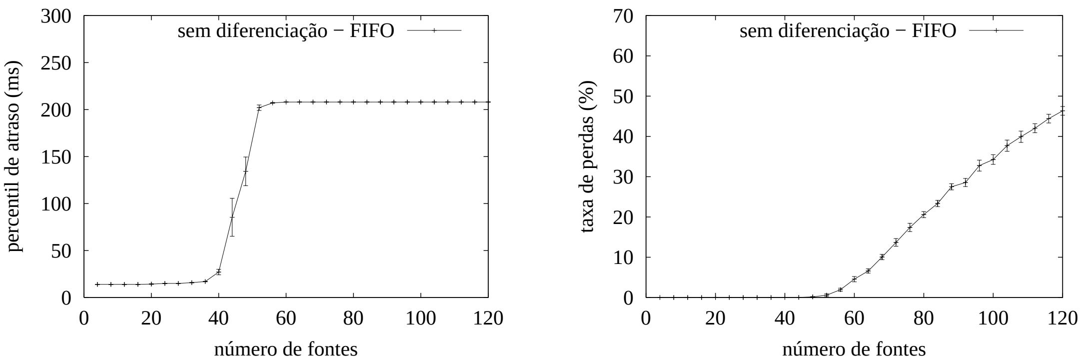
As fontes B0, conforme descrito na seção 4.1.2, representam nós móveis que já estavam acessando a rede móvel antes da situação de sobrecarga. Os novos nós que chegam, causando a sobrecarga, são pertencentes a sub-classe B1. Inicialmente, 40 fontes da sub-classe B0 encontram-se transmitindo na rede. Essas fontes geram em conjunto uma taxa média de 1,024 Mbps, representando 40,96% do enlace de gargalo. Este número permanece constante e as novas fontes que chegam são encaminhadas para a sub-classe B1. A característica de mobilidade modelada nesse caso foi a imprevisibilidade do movimento, que pode gerar as situações de sobrecarga.

Os esquemas simulados representam casos onde classes flexíveis são empregadas de forma a maximizar o número de usuários atendidos. Deseja-se observar as características do serviço oferecido por cada um dos esquemas. A alocação de recursos para as classes nos mecanismos SPP e PQ não é explicitamente controlável, como o que ocorre na arquitetura empregando o WRR. O escalonador PQ oferece garantias ao tráfego mais prioritário, chegando a disponibilizar toda a banda-passante do enlace de gargalo para ele, caso seja necessário. Isso faz com que o PQ garanta uma alocação de recursos adequada para servir as necessidades da sub-classe mais prioritária sempre que a carga gerada por essa for menor que a capacidade do enlace. Essa característica, entretanto, não permite o controle do serviço da sub-classe de baixa prioridade, como será mostrado nos testes realizados. O SPP, por sua vez, não realiza alocação de banda-passante para as sub-classes, mas emprega um gerenciamento de *buffer* de forma a diferenciar as sub-classes. Na arquitetura com WRR, os condicionadores de tráfego nas bordas dos domínios para cada sub-classe são configuráveis, permitindo que o agregado possua características de tráfego conhecidas. Além disso, o escalonador garante a alocação de fatias de serviço para cada sub-classe de tráfego. Dessa forma, utilizando-se o conjunto escalonador/condicionador para cada sub-classe, é possível que se garanta que o provisionamento se mantenha adequado dentro do domínio em questão. Assim, espera-se que os serviços oferecidos pela arquitetura baseada no WRR sejam altamente dependentes da configuração do mecanismo. No caso do PQ e do SPP, pode ser dito que a alocação de recursos (conseqüentemente, seu correto provisionamento) fica implícita ao mecanismo.

O que se deseja observar é a capacidade de cada mecanismo em respeitar a definição de cada classe de serviço, oferecendo sempre garantias de baixos atrasos. Dessa forma, os mecanismos são testados servindo classes de tempo real, sendo que uma das sub-classes permite degradação em um parâmetro de QoS para oferecer garantias à outros.

5.3.1 Avaliação

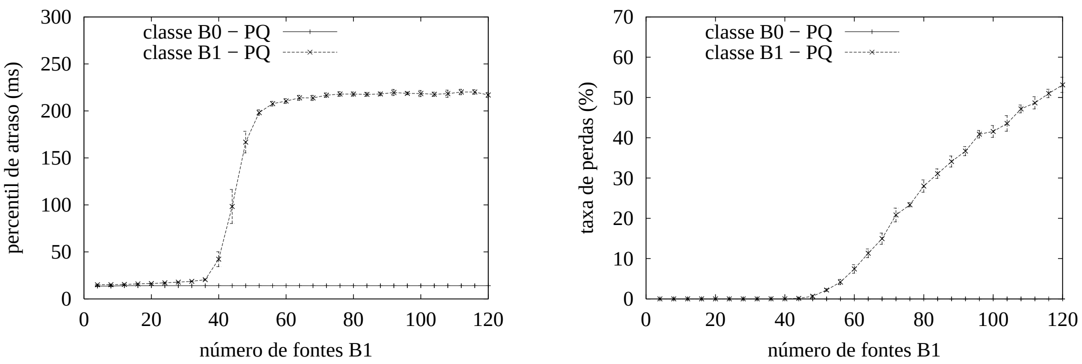
Inicialmente, a título de comparação, realizou-se uma simulação em que os fluxos não sofrem diferenciação alguma (figura 5.2). Nesse caso, todas as filas implementadas são FIFO do tipo *droptail*. Os resultados apresentados no gráfico 5.2(a) mostram o percentil-95 de atraso, em milisegundos, com o aumento do número de fontes B1. Nota-se que quando a carga das fontes se aproxima da capacidade do enlace de gargalo, ou seja, o número total de transmissores fica próximo das 97 fontes (40 fontes inicialmente transmitindo mais 57 que chegam depois), o atraso de pacotes cresce sem nenhum tipo de controle, até que se atinja o tempo necessário para se servir a fila completamente cheia. A partir desse instante, pacotes passam a ser descartados e a taxa de perdas passa a crescer (figura 5.2(b)).



(a) Percentil-95. (b) Taxa de perdas.

Figura 5.2: Fila única sem diferenciação. O número de fontes indicado não inclui as 40 inicialmente transmitindo.

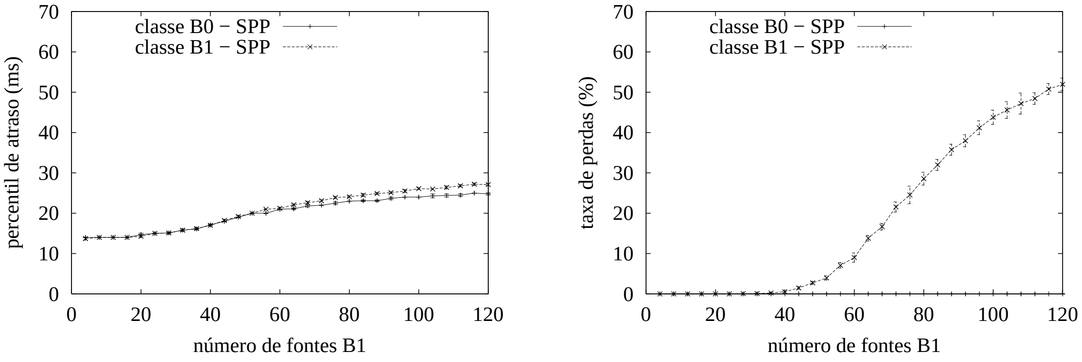
Em seguida, foram realizadas simulações utilizando-se os mecanismos propostos. A figura 5.3 apresenta os resultados da simulação para atraso e taxa de perdas obtidos na rede para o mecanismo PQ.



(a) Percentil-95. (b) Taxa de perdas.

Figura 5.3: Esquema com escalonador de prioridade PQ.

Observa-se, de acordo com a figura 5.3(a), que o atraso obtido para o tráfego B0 usando-se o mecanismo PQ mantém-se pequeno, conforme desejado. Em contrapartida, nessa implementação, a sub-classe B1 não oferece garantias quanto a atraso. Isso pode ser percebido pelo grande aumento do percentil-95 com a chegada de clientes dessa sub-classe. O atraso para a sub-classe B1 fica limitado quando sua fila fica completamente ocupada e o tempo de serviço para de crescer. Nesse instante a taxa de perdas da sub-classe de menor prioridade começa a apresentar um crescimento, conforme apresentado na figura 5.3(b). As perdas da sub-classe B0 ficaram muito baixas, em torno de zero, respeitando a especificação da sub-classe. A taxa de perdas crescente para a sub-classe B1 representa uma característica prevista da mesma. O comportamento de grande crescimento no atraso para a sub-classe menos prioritária, inviabiliza a utilização desse mecanismo em tráfego de tempo real[1].



(a) Percentil-95. (b) Taxa de perdas.

Figura 5.4: Esquema com gerenciamento de filas SPP.

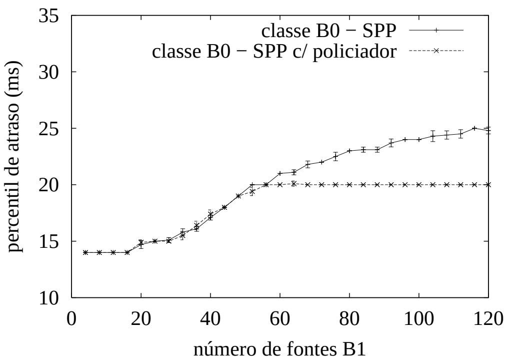
As simulações do mecanismo SPP demonstraram que sua curva de percentil, é mantida baixa (menor que 25 ms) para ambas as sub-classes. Sua taxa de perdas para a sub-classe B0 se mantém próxima de zero, como desejado. A sub-classe B1, entretanto, apresenta crescimento significativo na taxa de perdas, de acordo com a figura 5.4(b). Esse comportamento era esperado, estando previsto na especificação da classe.

A principal desvantagem observada no SPP é o crescimento do percentil de atraso da sub-classe B0 com o aumento do número total de fontes, fato que é notado pela

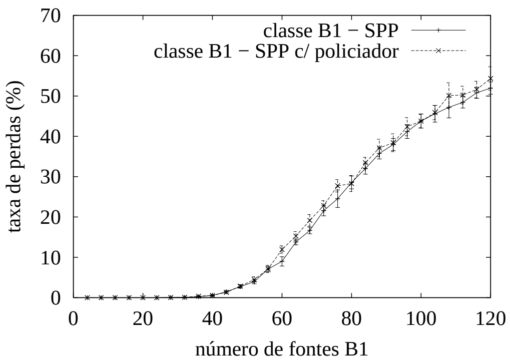
[1]Em telefonia, por exemplo, atrasos totais de até 150 ms são aceitáveis.

derivada positiva da curva do percentil. Isso mostra que, em termos de atraso, o tráfego B0 é influenciado pela chegada de fontes B1, o que não é desejável. O fato do atraso no tráfego da sub-classe B0 crescer de acordo com o aumento das fontes de sub-classe B1 demonstra também que o isolamento entre as sub-classes não é total. Esse aumento ocorre porque no SPP ambas as sub-classes ocupam a mesma fila. Assim, o crescimento na ocupação dessa fila traz aumentos no atraso para as duas sub-classes. Uma forma de se estabelecer um limite para a interferência causada pelo tráfego B1 no tráfego B0 é através da utilização de um mecanismo de policiamento de tráfego para o primeiro. Um *token bucket* [13] com *bucket* de tamanho igual a um pacote foi escolhido para exercer o policiamento de tráfego, descartando os pacotes da sub-classe B1 quando seu agregado ultrapassar uma taxa limite. Dessa maneira, o atraso global é mantido baixo, uma vez que as filas não são mantidas excessivamente cheias. Nesse caso, uma opção pela diminuição do atraso, em detrimento da taxa de perdas, foi realizada.

O policiamento foi realizado no nó de borda número 4, mantendo a taxa do agregado de fontes B1 limitada. A taxa limite foi escolhida de modo a permitir uma taxa média equivalente àquela gerada por 60 fontes. Quando a taxa média ultrapassa esse valor, pacotes passam a ser descartados. Essa simulação é apresentada na figura 5.5(a), onde é realizada uma comparação do SPP antes e depois da utilização do policiador.



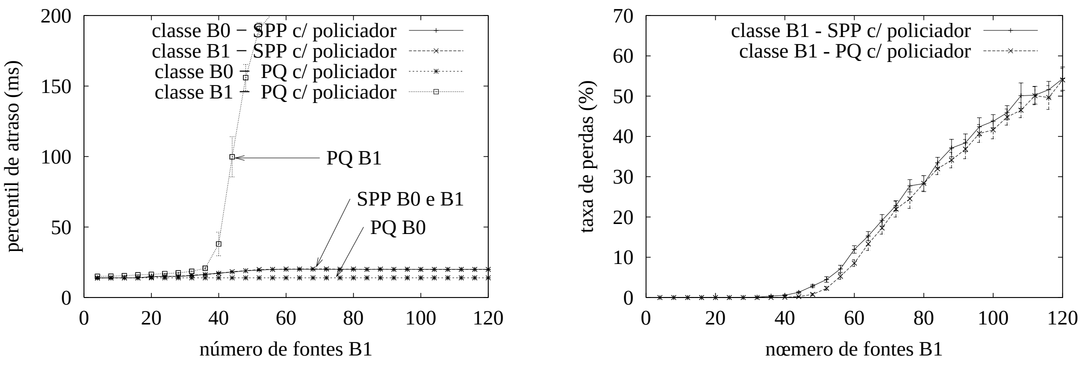
(a) Percentil-95.



(b) Taxa de perdas.

Figura 5.5: Efeito da inclusão do policiador na arquitetura de SPP.

A curva referente à sub-classe mais prioritária implementada através do SPP cresce de forma aproximadamente linear com o aumento do número de fontes. A partir da taxa equivalente a 60 fontes, entretanto, a curva passa a se manter constante, próxima à 20 ms. A taxa de perdas para o SPP, observada na figura 5.5(b), é ligeiramente superior que a observada no caso anterior, sem policiador. Esse fato era esperado, uma vez que o policiador descarta o tráfego excedente. As perdas para a sub-classe B1 do SPP ocorrem agora tanto no núcleo da rede, através de perdas na fila do nó, como também na borda, através do policiador.



(a) Percentil-95.

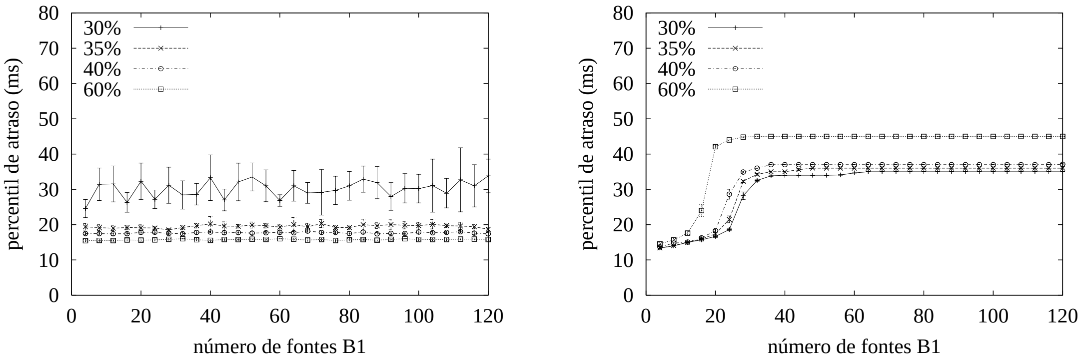
(b) Taxa de perdas.

Figura 5.6: Comparando o efeito do policiador para SPP e PQ.

Uma nova simulação foi realizada para garantir uma comparação justa entre os mecanismos SPP e PQ. Para tanto, incluiu-se o policiador para o tráfego B1 no esquema do PQ. Como observado na figura 5.6(a), não houve mudanças significativas. A alta taxa de crescimento na curva de percentil de atraso do tráfego B0, para o esquema com duas filas, se manteve, mesmo com a limitação de taxa imposta pelo policiador. Esse fato ocorre por causa de características particulares do escalonador de prioridade, que não serve tráfego B1 enquanto houver pacotes B0 na fila prioritária. A taxa de perdas do tráfego B1 para o esquema de duas filas sofre um ligeiro aumento, causado pelo policiador.

As simulações da arquitetura baseada no escalonador WRR incluem um elemento que não estava presente nas anteriores, um suavizador de tráfego para cada sub-classe. Esse elemento é empregado para garantir que ambas as sub-classes respeitem a definição de serviço do EF, conforme discutido na seção 4.2.2. Para isso é necessário que a taxa de suavização de cada sub-classe seja menor que a respectiva taxa de serviço, configurada no escalonador WRR. O suavizador utilizado foi um *leaky bucket* [13] que foi empregado nos roteadores de ingresso da Rede de Núcleo. O tamanho do *buffer* do suavizador para o tráfego B0, mais prioritário, é grande o suficiente para evitar descartes de pacotes. O provisionamento dessa sub-classe é realizado da seguinte forma: o suavizador está configurado de modo a permitir a transmissão adequada de 40 fontes simultaneamente. Para que essas transmissões sejam adequadas a taxa individual de cada fonte, utilizada como base para o cálculo da taxa de suavização, foi escolhida com valores 30%, 35%, 40% e 60% maiores que a taxa média de cada micro-fluxo. Assim, para o cálculo utilizando-se uma margem de 35% sobre a taxa média (a taxa média é igual a 25600 bps por micro-fluxo) obtém-se uma taxa individual de 34560 bps. A taxa de suavização para 40 fontes é então 1382400 bps. O escalonador foi configurado para fornecer à sub-classe B0 uma fatia de serviço com taxa 10% superior à taxa de suavização do agregado. No caso do exemplo citado, essa fatia corresponde a 0.608256. A fatia restante destina-se à sub-classe menos prioritária B1. O suavizador para a sub-classe B1 foi configurado para uma taxa igual a 90% da porção reservada à sub-classe menos prioritária pelo WRR. No caso do exemplo, a taxa de suavização da sub-classe B1 seria igual a 979360 bps, correspondendo à transmissão simultânea de aproximadamente 28 fontes, considerando-se a folga de 35% (ou 38 fontes, considerando-se a taxa média).

Observa-se que a variação da taxa de suavização e conseqüentemente da fatia do enlace reservada pelo WRR para a sub-classe mais prioritária leva a diferentes resultados em termos de atraso para ambas as sub-classes e taxa de perdas para a sub-classe B1. Ao se aumentar as reservas para a sub-classe mais prioritária, o atraso dessa sub-classe diminui, conforme apresentado na figura 5.7(a). Em contrapartida, o atraso e as taxas de perdas para a sub-classe de baixa prioridade aumentam (figuras 5.7(b) e 5.8). Nota-se grande sensibilidade do atraso na sub-classe B0 à configuração da taxa de suavização e da alocação da fatia de serviço no WRR. Ao se modificar a taxa de suavização de 30% para 35% superior a taxa média, é possível verificar-se uma grande diferença na curva de percentil de atraso.



(a) Classe B0. (b) Classe B1.

Figura 5.7: Percentil-95 de atraso para as sub-classes do WRR com diferentes configurações do suavizador B0.

Observa-se que o atraso para a sub-classe mais prioritária permanece aproximadamente constante, mesmo com o aumento no número de transmissores da sub-classe B1. Esse fato demonstra que essa arquitetura oferece boa independência entre as sub-classes. Isso se dá por que o provisionamento de recursos é realizado de forma fixa para cada uma das sub-classes. Conforme visto na figura 5.7(b), na sub-classe de menor prioridade B1, ocorre um crescimento do atraso quando a quantidade de fontes transmitindo se aproxima da taxa de suavização para essa sub-classe, ou seja, quando a alocação estática dos recursos passa a ser inadequada. O atraso pode ser mantido baixo e limitado devido aos descartes que passam a ocorrer no suavizador de tráfego, que não permitem que o *buffer* do nó de entrada do enlace de gargalo fique sobrecarregado.

Os gráficos de taxas de perdas da sub-classe de baixa prioridade B1 para os mecanismos simulados foram sumarizados na figura 5.9 de forma a facilitar comparações. A taxa de perdas oferecida à sub-classe B1 pelo mecanismo PQ fica ligeiramente abaixo da observada com o SPP, no entanto, o PQ não oferece garantias de baixos atrasos para a sub-classe de menor prioridade B1. Conforme discutido anteriormente, a taxa de perdas para o WRR depende fortemente das configurações escolhidas. Para uma comparação, foi escolhida a configuração com taxa de suavização 35% superior a taxa média para a sub-classe B0. Nesse caso, o atraso

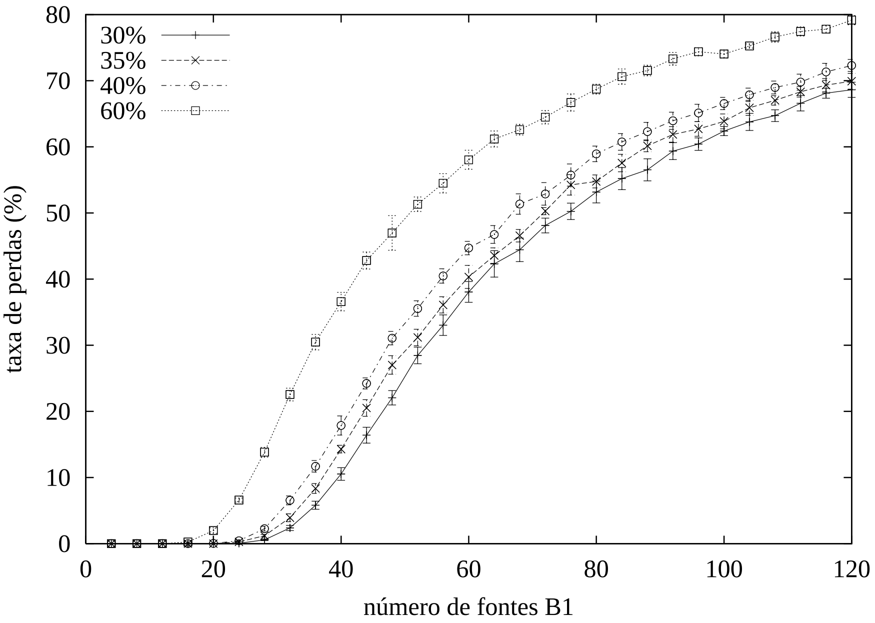


Figura 5.8: Taxa de perdas para diferentes configurações do suavizador B0 na arquitetura com WRR.

da sub-classe B0 fica abaixo do observado na sub-classe B1 e, ao mesmo tempo, o serviço da sub-classe B1 sofre pouca degradação nos períodos de sobrecarga. Nessa configuração, a taxa de perdas obtida para o WRR ficou acima das encontradas para os outros mecanismos. A menor taxa de perdas foi observada para o mecanismo sem diferenciação de classes. Isso ocorre porque as filas dos roteadores permaneceram com uma maior ocupação, não ocorrendo nenhum tipo de limitação de tráfego para nenhuma das sub-classes propostas. É importante notar que nesse caso as perdas estão distribuídas igualmente entre todas as fontes, não existindo o conceito das sub-classes. Assim, essa opção não fornece nenhum tipo de garantia ou diferenciação de QoS, sendo apresentada como referência do caso de utilização de uma rede somente com serviço *best effort*.

Observando-se o gráfico de taxas de perdas, pode ser realizada uma comparação da quantidade de usuários da sub-classe menos prioritária que pode ser admitida em cada um dos esquemas propostos. Para que se mantenha a satisfação do usuário em 3 (numa escala MOS, de acordo com a tabela 2.1), uma taxa de perdas em torno de 20% deve ser mantida. A tabela 5.2 apresenta o numero de fontes admitidas para a

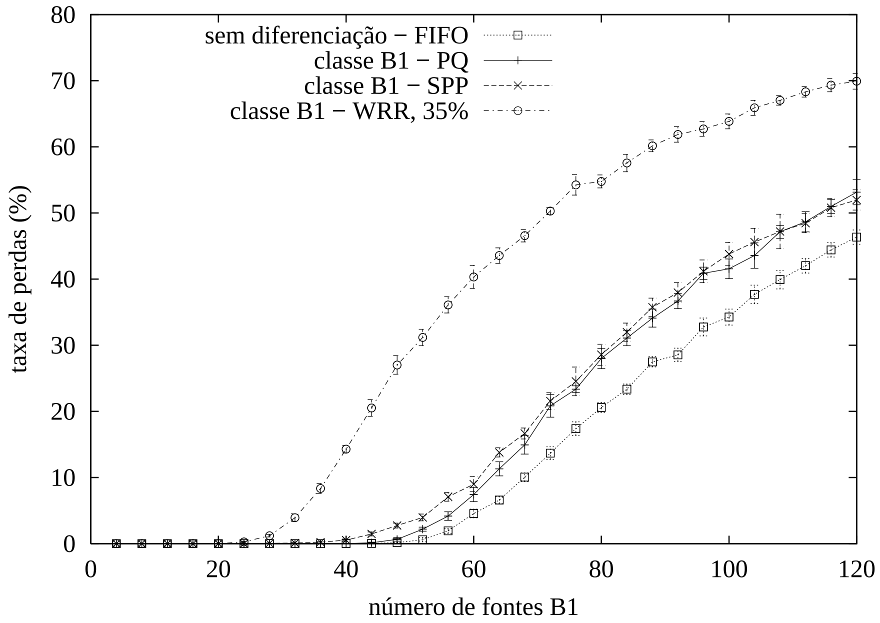


Figura 5.9: Taxa de perdas para os três mecanismos.

Tabela 5.2: Número de fontes admitidas e percentil-95 de atraso para taxas de perdas em torno de 20%.

Arquitetura	Fontes B1 admitidas	Atraso B0 (ms)	Atraso B1 (ms)
PQ	72	14	216
SPP com policiador	68	20	20
WRR 35%	44	20	36

sub-classe B1 e o percentil-95 de atraso medido em ambas as sub-classes para essa taxa de perdas.

As simulações mostraram que o mecanismo PQ, apesar de oferecer os melhores resultados de atraso e perdas para a sub-classe B0, não oferece nenhum tipo de garantia em termos de atraso para a sub-classe B1. Assim, tal mecanismo não é adequado para um esquema de suporte e diferenciação das sub-classes B0 e B1. Isso porque a sub-classe B1 tem como objetivo dar suporte a aplicações multimídia de tempo real. Desse modo, mesmo sendo de menor prioridade com relação a perdas, ela deve receber serviço de baixos atrasos, de modo similar ao da sub-classe B0. O mecanismo de SPP, quando aplicado em conjunto com um policiador do tráfego

B1, conseguiu fornecer um serviço com atraso baixo e limitado para ambas as sub-classes. Além disso, as condições de taxas de perda próximas de zero para o tráfego B0 foram satisfeitas. A taxa de perdas para a sub-classe B1, entretanto, ficou levemente superior à encontrada com o mecanismo PQ, levando a um menor número de usuários transmitindo simultaneamente, em relação ao PQ, para a taxa de perdas de 20% . A arquitetura baseada em WRR mostrou-se fortemente dependente de configuração correta dos seus componentes, o que pode ser difícil em uma rede que precisa lidar com mobilidade e conseqüentemente altas variações em suas necessidades de recursos. Assim, para a taxa de perdas de 20% o número de fontes B1 transmitindo fica bem abaixo daquele observado no PQ e no SPP. Nos mecanismos WRR e SPP ocorre uma solução de compromisso que troca taxas de perdas por atraso para a sub-classe menos prioritária B1.

5.4 Avaliação do MBAC

As simulações dos mecanismos de suporte para a classe B, apresentadas na seção anterior, demonstram que a sub-classe de menor prioridade B1 sofre um aumento na taxa de perdas durante as situações de sobrecarga. Esse comportamento estava previsto na descrição da classe, sendo uma solução de compromisso, entre taxa de perdas e atraso. O nível máximo de degradação dos serviços para a sub-classe menos prioritária, entretanto, deve ser controlável para que não ocorram situações onde a rede fique inutilizável. Para isso, esquemas de controle de admissão podem ser empregados. No caso das Redes de Núcleo de sistemas celulares 3G existem diversas características que fazem com que a utilização de controle de admissão baseado em medidas (MBAC) seja vantajosa, conforme discutido na seção 4.3.1.

Assim, será realizada nessa seção uma avaliação do comportamento de um mecanismo de MBAC em um ambiente de mobilidade. O mecanismo de MBAC escolhido foi o de Soma Medida, que foi apresentado na seção 4.4 para dar suporte a classe B.

Ao se considerar a utilização do mecanismo em Redes de Núcleo de sistemas celulares torna-se importante estudar seu comportamento em diferentes níveis de mobilidade. Dessa forma, nas simulações apresentadas a mobilidade estará modelada através de taxas de chagada e de saída de fluxos de dados aos nós da rede. Assim, a topologia simulada (figura 5.1) representa uma Rede de Núcleo de um sistema móvel, onde usuários realizam *handover*. Os nós onde os usuários móveis se conectam representam um roteador de borda da Rede de Núcleo (na arquitetura UMTS, um 3G-SGSN). Dessa forma, foram realizadas simulações onde a mobilidade foi modelada a partir de diferentes taxas de chegada de fontes e diferentes tempos de duração de chamadas, que representam o tempo de permanência de uma fonte em uma dada região, transmitindo através de um determinado roteador de borda. Essa modelagem foi realizada observando-se que quando um nó possui maior velocidade de movimentação, ele permanece durante uma menor quantidade de tempo transmitindo em uma dada região, quando comparado à um nó mais lento. Essa característica foi mapeada através de um maior tempo de duração de chamadas para as fontes com menor mobilidade. Da mesma maneira, um ambiente com nós se movimentando com maior velocidade apresenta uma maior quantidade de *handovers* em seu sistema. Isso se reflete em uma maior taxa de chegada de fontes em uma dada região. Assim, o menor tempo entre chegadas é observado em ambientes de muita mobilidade.

Em todos os diferentes cenários de mobilidade, o número médio de fontes tentando transmitir simultaneamente (ou seja, a carga média a qual o MBAC foi submetido) foi mantido constante de forma a permitir uma comparação coerente. As medições somente foram consideradas quando o número de fontes tentando acessar a rede fosse constante, ou seja, atingisse o estado estacionário. O intervalo entre as chegadas de fontes segue uma variável aleatória com distribuição exponencial. O mesmo ocorre com o tempo de duração das chamadas. Desta forma, o número médio de fontes ρ tentando transmitir simultaneamente pode ser calculada a partir da equação $\rho = \eta/\lambda$ [66]; onde η é a média do tempo de duração da chamada e λ é a média do intervalo entre chegadas.

Assim, o número médio de fontes tentando acessar a rede foi escolhido acima da capacidade de seu enlace de gargalo em todos os níveis de mobilidade testados.

Tabela 5.3: Parâmetros dos diferentes níveis de mobilidade simulados.

Mobilidade	Tempo entre chegadas - λ (seg)	Duração das chamadas - η(seg)
Muita	0,1	30
Média	0,5	150
Pouca	0,9	270
Nenhuma	0,9	–

Nesse cenário, o mecanismo de MBAC será testado em uma situação de limite da capacidade da rede. O número médio de fontes tentando acessar a rede foi então escolhido igual a 300 fontes. Os parâmetros de configuração de mobilidade para cada um dos cenários encontram-se descritos na tabela 5.3. O cenário "nenhuma mobilidade" foi configurado para representar um caso onde ocorrem apenas chegadas de fontes à rede, sendo que as transmissões efetuadas pelas fontes aceitas terminam somente ao final da simulação.

5.4.1 Sintonia do MBAC

Conforme discutido na seção 4.4, o desempenho do mecanismo de MBAC Soma Medida depende da correta escolha de parâmetros, tanto do estimador quanto do critério de admissão. No caso do mecanismo de Soma Medida em conjunto com o estimador Janela de Tempo, os parâmetros de ajuste de desempenho são:

- nível de utilização da rede, representado por υ;
- tempo de duração da janela de medição T;
- tempo de duração da amostra S.

De acordo com o discutido na seção 4.4.1, a escolha desses parâmetros faz com que o controle de admissão seja mais ou menos conservador, ou seja, tenha menor ou maior chance de permitir a ocorrência de degradação dos serviços com a admissão de novos fluxos. É importante notar que uma maior chance de degradação implica em um maior número de fluxos admitidos.

A literatura indica que a escolha do tempo de duração da janela de medição (parâmetro T) é o fator que tem maior influência no desempenho do MBAC [83]. Por esse motivo buscou-se observar o comportamento do controle de admissão face a variação desse parâmetro. Para isso, manteve-se constante a relação entre o tamanho da janela de medição e o tempo de amostragem S, ou seja, o número de medições realizadas em cada janela. Seguiu-se a recomendação de manter a relação $T/S = 10$. Assim, o tempo de amostragem S foi variado em conjunto com o tempo da janela T. Isso permite a obtenção do ponto de melhor sintonia do MBAC, considerando-se a influência do conjunto de parâmetros sobre seu desempenho. O parâmetro v que indica a utilização máxima da rede foi mantido constante em 90%.

Assim, variou-se o tamanho da janela T e observou-se o número médio de fontes admitidas na rede. Nessa simulação não foi utilizado nenhum mecanismo de diferenciação de classes, empregando-se uma única classe e utilizando-se filas FIFO na topologia da figura 5.1.

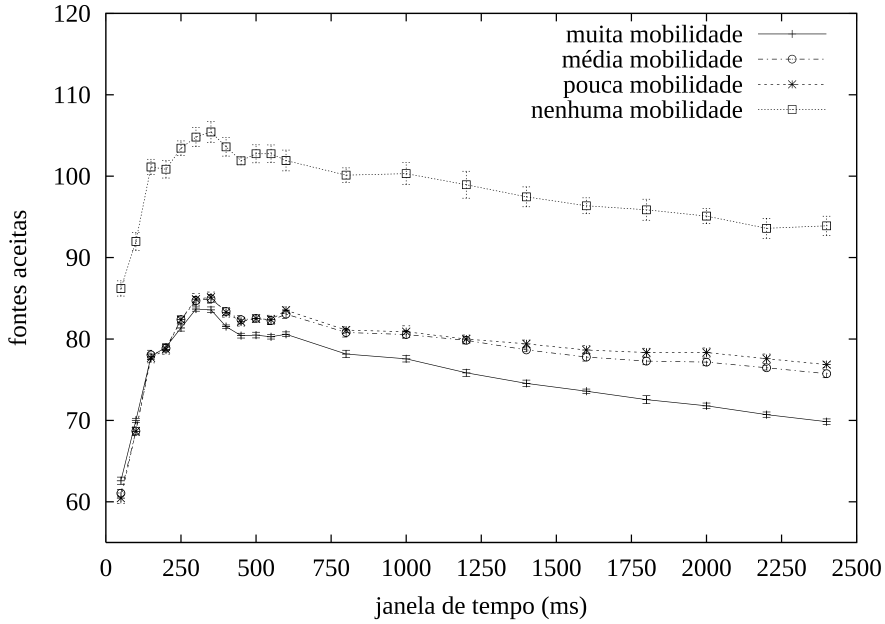


Figura 5.10: Número médio de fontes admitidas para diferentes durações da janela de tempo.

A partir das simulações observa-se que o comportamento do MBAC não apresenta grande variação em função do nível de mobilidade, de acordo com a figura 5.10.

Essa figura traz o número médio de fontes admitidas para diferentes tamanhos de janela de tempo T. São estudados os quatro níveis de mobilidade apresentados na seção anterior. O caso em que não há mobilidade, onde as fontes somente chegam para disputar o meio e onde as transmissões não terminam, representa um limite superior na quantidade de fontes admitidas. Isso se dá porque nesse cenário a carga a qual o MBAC é submetida é crescente e muito superior a dos demais cenários. Nas outras situações de mobilidade, quanto mais dinâmico o cenário, pior é o comportamento do MBAC em termos de número médio de fontes admitidas. Entretanto, é interessante notar que todos os níveis de mobilidade conseguem admitir o maior número de fontes em torno de um mesmo tamanho de janela, próximo a 350 ms. Nesse ponto a diferença de desempenho do mecanismo de controle de admissão em termos do número de fontes admitidas para os diversos níveis de mobilidade, excluindo-se o cenário "nenhuma mobilidade", é muito pequena.

O ponto de máxima admissão representa a configuração menos conservadora do MBAC. Nesse ponto pode ocorrer um certo nível de degradação dos serviços. Esse nível de degradação, entretanto, pode ser controlado pelo parâmetro do MBAC que indica a utilização da rede, o parâmetro v. Assim, o nível de serviço oferecido às sub-classes pode ser controlado através da correta configuração do MBAC. De acordo com a proposta de serviços flexíveis, é interessante que haja níveis controláveis de degradação em alguns parâmetros de QoS, desde que em contrapartida ocorram melhorias em outros. No caso da classe B, aumentos controlados nas taxas de perdas podem ocorrer para que o número de usuários admitidos possa ser aumentado e conseqüentemente a probabilidade de desconexão diminuída. Os mecanismos de suporte a classe B irão atuar no caso de eventuais sobrecargas, oferecendo as garantias à sub-classe mais prioritária B0, e mantendo os baixos atrasos para ambas as sub-classes.

5.4.2 MBAC para classe B

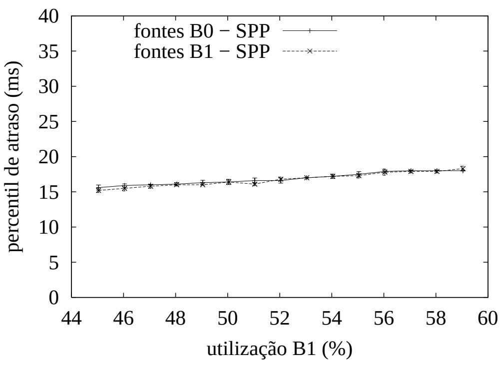
A utilização do MBAC na classe B deve permitir o controle do nível de degradação do desempenho das sub-classes em situações de sobrecarga. Dessa forma, é importante que se avalie esses diferentes níveis de degradação que podem ocorrer dependendo da configuração do MBAC.

O controle de admissão estudado foi então implementado em dois níveis, um para cada sub-classe, seguindo-se o que foi proposto na seção 4.5. Através do parâmetro υ, de configuração da utilização da rede por cada sub-classe, é possível realizar um controle indireto do nível de degradação obtido e também do número de clientes admitidos em cada sub-classe. No caso da topologia estudada, o parâmetro υ representa a utilização no enlace de gargalo. Assim, foram realizadas simulações onde o parâmetro υ foi variado. A partir dessas simulações é possível verificar o comportamento do MBAC em conjunto com os mecanismos propostos para a implementação da classe B e a viabilidade da utilização do MBAC para o controle do nível de degradação da mesma.

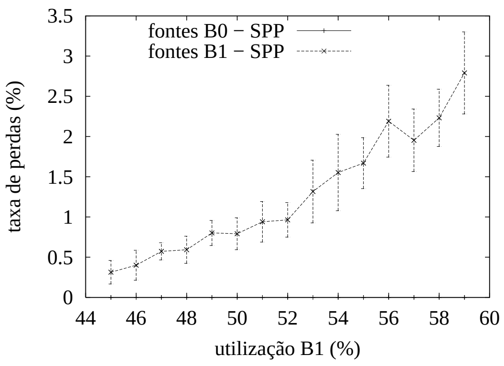
O MBAC foi configurado para aceitar o número máximo de usuários, baseando-se nos resultados da seção 5.4.1, uma vez que a probabilidade de desconexão deve ser minimizada. Assim, o tempo de duração da janela de medição T foi escolhido igual a 350 ms e a razão T/S foi mantida igual a 10. O parâmetro de utilização da sub-classe mais prioritária B0 foi configurado para permitir que essa classe impusesse uma carga a rede equivalente a 40,96% do enlace de gargalo. Essa configuração deve permitir aproximadamente 40 fontes dessa sub-classe transmitindo simultaneamente na rede. O cenário de mobilidade escolhido foi o de "muita mobilidade", apresentado na tabela 5.3.

As simulações demonstraram que, nas três arquiteturas simuladas, SPP sem policiador, WRR 35% e PQ, o número de fontes da sub-classe B0 manteve-se próximo a 40, o que está de acordo com o desejado pela configuração realizada com o MBAC. O número de clientes da sub-classe B1 apresenta crescimento de acordo com o aumento de υ, também de acordo com o esperado.

Medidas de desempenho para atraso e taxa de perdas obtidas a partir das simulações de MBAC em conjunto com a arquitetura baseada no SPP estão apresentadas na figura 5.11. Nota-se que, similarmente ao observado na seção 5.3.1, o aumento na utilização planejada para a sub-classe B1 influencia negativamente o atraso de am-



(a) Percentil-95.



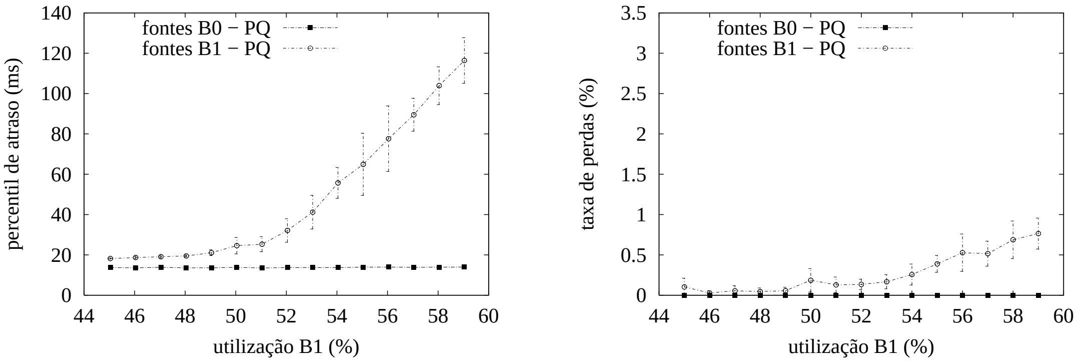
(b) Taxa de perdas.

Figura 5.11: Variação de υ para classe B1 com SPP.

bas as sub-classes (figura 5.11(a)). Não há sentido em se aplicar um policiador nesse caso, pois o controle de admissão está sendo realizado em nível de chamada. Na realidade, o controle de admissão por chamada irá justamente evitar a degradação excessiva que pode ocorrer nas situações de sobrecarga, quando apenas o policiamento é empregado para controle do volume de tráfego. Ao se utilizar o policiador para uma sub-classe, todas as conexões dessa são degradadas. Caso o nível de degradação não seja controlável, todas as conexões podem ser inutilizadas. Quando se utiliza um controle de admissão por chamadas, como o MBAC, é possível se controlar o nível de degradação máximo permitido. Caso se deseje obter um determinado percentil de atraso ou uma taxa de perdas para as sub-classes, o parâmetro υ para a sub-classe menos prioritária deve ser configurado adequadamente no MBAC.

O MBAC demonstrou-se eficiente para manutenção de pequenas taxas de perdas da sub-classe B1 (inferiores a 3,5%) para todas os valores de υ simulados. As taxas de perda da sub-classe de maior prioridade B0 foram mantidas nulas, conforme observado no gráfico da figura 5.11(b).

Os resultados das simulações para a arquitetura baseada em PQ encontram-se apresentadas na figura 5.12. Nota-se que nessa arquitetura, mesmo com a utilização do MBAC, a manutenção de pequenos atrasos para a sub-classe B1 depende de baixos valores de υ, que por sua vez implica na presença de poucas fontes dessa classe transmitindo na rede. As taxas de perda observadas para a sub-classe B1 são



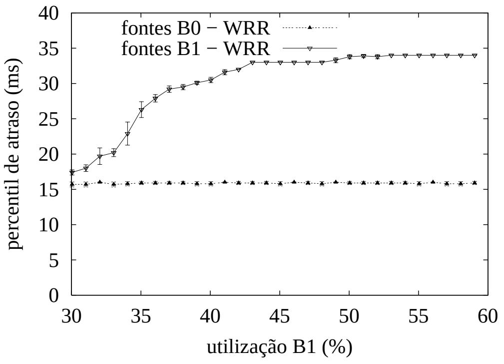
(a) Percentil-95. (b) Taxa de perdas.

Figura 5.12: Variação de v para classe B1 com PQ.

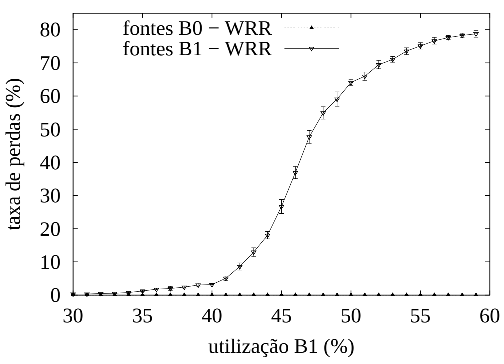
inferiores às obtidas com a arquitetura baseada no SPP. A sub-classe B0 obtém os níveis de desempenho desejados, com baixos atrasos e de taxa de perdas nula.

A configuração utilizada nas simulações com o WRR foram semelhantes às da seção 5.3.1, para taxa de suavização 35% superior a taxa média para o agregado B0. Assim a fatia de serviço no escalonador WRR alocada para a sub-classe B0 foi 10% superior à taxa de suavização calculada com folga de 35%. Isso resultou em um peso de aproximadamente 61% para a sub-classe B0 no WRR. O peso da sub-classe B1 foi então os 39% restantes. Os suavizadores das sub-classes B0 e B1 foram configurados para fornecer uma taxa igual a fatia alocada pelo escalonador para cada sub-classe, sem a folga de 10% empregada nas simulações da seção 5.3.1. Essa configuração foi escolhida porque o MBAC deve limitar o tráfego garantindo o correto provisionamento dos recursos. A variação de v para a sub-classe B1 irá permitir a configuração adequada do MBAC, mostrando quais valores desse parâmetro levam a obtenção do serviço especificado.

Gráficos de atraso e perdas obtidos nas simulações para o WRR são apresentados na figura 5.13. Observa-se que o atraso para a sub-classe B0 manteve-se baixo, graças ao correto provisionamento de recursos para essa classe. A sub-classe B1, entretanto, demonstra um crescimento do atraso quando a utilização dessa classe se aproxima de 40%. Nota-se que a partir desse momento a taxa de perdas apresenta uma acentuada elevação. Quando o valor v da utilização permitida ultrapassa a taxa de suavização,



(a) Percentil-95.



(b) Taxa de perdas.

Figura 5.13: Variação de υ para classe B1 com WRR.

Tabela 5.4: Valor de υ, número de fontes admitidas e percentil-95 de atraso para taxas de perdas em torno de 0,8%.

Arquitetura	υ	Fontes admitidas	Atraso B0 (ms)	Atraso B1 (ms)
PQ	59%	39 B0; 49 B1	14	116
SPP	50%	40 B0; 43 B1	16	16
WRR 35%	34%	40 B0; 29 B1	15	23

o MBAC passa a funcionar de forma inadequada. Isso ocorre porque mesmo que o número de clientes transmitindo na rede gere uma carga muito superior a desejada pelo MBAC, a taxa observada no enlace de gargalo (enlace onde o MBAC realiza as medidas, representando o interior da Rede de Núcleo) fica limitada à taxa de suavização. Isso faz com que sejam aceitos mais fluxos do que o máximo possível para a manutenção de um serviço adequado.

As simulações demonstraram que o MBAC pode ser aplicado em conjunto com os mecanismos propostos na seção 4.2 para que se limite a degradação dos serviços oferecidos pela sub-classe B1. A tabela 5.4 apresenta os valores do parâmetro υ para a sub-classe B1 comparando-se níveis de serviço equivalentes em termos de taxas de perdas nas três implementações da classe[2]. São apresentados também os valores

[2] O valor da taxa de perdas escolhido para comparação, de 0,8%, foi o maior observado para a arquitetura baseada em PQ.

do percentil-95 para atraso em ambas as sub-classes. Observa-se que apesar de permitir um maior número de fontes admitidas, o que leva a uma maior utilização, a arquitetura baseada no PQ não oferece garantias de atraso para a sub-classe B1. A utilização em conjunto do SPP com o MBAC permitiu os serviços de baixos atrasos para ambas as sub-classes, com taxa de perdas controlável para a sub-classe B1. A interação do MBAC com o WRR mostrou-se bastante complexa. Deve se ter muito cuidado na configuração dos suavizadores, respeitando-se essa configuração ao realizar a escolha dos valores de v para ambas as sub-classes, uma vez que a suavização força artificialmente a taxa de cada agregado no enlace de gargalo. Assim, o MBAC atua de maneira inadequada quando a utilização configurada para cada sub-classe se aproxima da respectiva taxa de suavização. A configuração correta do conjunto de mecanismos nesse caso torna-se ainda mais importante, dificultando sua utilização em redes que lidam com mobilidade. Além disso, a arquitetura baseada em WRR foi a que possibilitou a menor quantidade de fontes admitidas para um mesmo nível de serviço.

Esse capítulo apresentou as simulações de avaliação de desempenho e comparação de alguns dos mecanismos propostos ao longo do texto. O capítulo seguinte traz as conclusões e considerações finais sobre o trabalho.

Capítulo 6

Conclusões

ESTE trabalho apresentou um estudo de qualidade de serviço em ambientes móveis, concentrando-se nas questões relativas à Rede de Núcleo de sistemas celulares de terceira geração. As Redes de Núcleo são responsáveis pelo transporte de todo o tráfego trocado entre os terminais móveis e as redes fixas, sendo peça fundamental na construção de uma arquitetura de serviços com QoS fim-a-fim. Essas redes devem ser capazes de oferecer suporte a serviços com requisitos rígidos em alguns parâmetros de QoS, como o exigido pelas aplicações de tempo real de voz e vídeo interativos. As soluções de QoS nessas redes devem lidar não apenas com as dificuldades tradicionais das redes IP, mas também com novos fatores relacionadas com a mobilidade, como a imprevisibilidade do movimento dos nós. Assim, as Redes de Núcleo podem sofrer grandes variações em seus níveis de utilização, o que dificulta um provisionamento adequado de QoS que simultaneamente permita altos níveis de utilização dos recursos. Altos níveis de utilização representam ainda um maior número de usuários atendidos, o que otimiza um parâmetro de QoS importante em ambientes de mobilidade: a probabilidade de desconexão.

Observou-se que as principais soluções de QoS propostas para a Internet nem sempre irão possibilitar o atendimento dos requisitos das Redes de Núcleo. Elas representam paradigmas opostos em termos de propriedade escalar e acurácia. O fator de mobilidade traz dificuldades para ambas as soluções. A movimentação dos nós exigiria freqüente renegociação de recursos no caso do IntServ e também tornaria difícil a realização de um provisionamento estático adequado, no caso da utilização do DiffServ.

Foi proposta então uma estrutura de diferenciação de serviços voltada para o ambiente de mobilidade, baseando-se na criação de classes de serviço flexíveis que permitem a degradação de alguns parâmetros de QoS para a garantia de outros. Essa flexibilidade foi considerada a partir das características de cada tipo de aplicação, aproveitando-se a robustez delas em relação a alguns parâmetros de QoS.

Em seguida, foram propostos diferentes mecanismos para a implementação de uma das classes sugeridas: a classe B, de tempo real em mobilidade. Três arquiteturas foram estudadas, duas baseadas em escalonadores tradicionalmente utilizados para suporte a QoS na Internet e a terceira baseada na política protetora de gerenciamento de *buffer* SPP, que apresentava características importantes para a classe em questão. Mecanismos de controle de tráfego foram empregados em conjunto com os escalonadores (PQ e WRR) e com o SPP. Assim, o policiamento de pacotes e controle de admissão por chamadas compuseram as arquiteturas propostas para implementação da classe em questão. Avaliações baseadas em simulações foram então realizadas, permitindo a comparação dessas diferentes arquiteturas.

A partir das simulações, observou-se que a arquitetura de QoS baseada na política protetora SPP apresentou mais características interessantes para o suporte à classe B do que os esquemas baseados no PQ e no WRR. Todas as políticas ofereceram bom isolamento entre classes em termos de perdas. A classe prioritária B0 manteve-se protegida, não apresentando perdas significativas.

A arquitetura baseada em PQ ofereceu os melhores resultados de atraso para classe prioritária B0, oferecendo valores cerca de 20% menores que os demais mecanismos para um mesmo nível de degradação. Além disso, essa arquitetura permitiu um maior número de usuários simultâneos da sub-classe B1, quando comparada às outras implementações (64% a mais que o WRR e 6% a mais que o SPP, todos com mesma taxa de perdas). O PQ, entretanto, não ofereceu garantias de atraso à sub-classe B1. Por sua vez, o mecanismo SPP manteve o atraso baixo para ambas as classes, apesar de sua curva de percentil-95 de atraso aumentar de forma aproximadamente linear, aparentemente sem garantias de limitação. A partir dessa característica, verificou-se a necessidade de policiamento de tráfego para garantir um valor máximo de atraso das sub-classes, independentemente do volume de tráfego B1. O mecanismo empregado foi de controle de admissão em nível de pacotes, através do algoritmo de *token bucket*. Após o emprego do policiamento do tráfego em excesso na arquitetura baseada em SPP, o atraso observado ficou igual ao medido na arquitetura do WRR para a sub-classe B0 (para uma mesma taxa de perdas na sub-classe B1). Ao se comparar os resultados para sub-classe B1, observa-se que o percentil-95 de atraso do SPP fica cerca de 44% abaixo do observado no WRR, para níveis equivalentes de degradação. O desempenho do SPP também mostra-se superior ao do WRR quando se compara o número de usuários da sub-classe de menor prioridade aceitos. Nesse caso, a arquitetura baseada no SPP permite cerca de 55% de usuários simultâneos a mais, em relação à configuração do WRR simulada. Os resultados da avaliação para o WRR mostraram que apesar desse mecanismo ser capaz de lidar com tráfego em tempo real ele é muito dependente da correta configuração de elementos da arquitetura, que inclui suavizadores para cada classe. De uma forma geral, tanto o WRR como o SPP realizam soluções de compromisso entre atraso e taxa de perdas para a sub-classe de menor prioridade. Dentre as arquiteturas simuladas, as baseadas no SPP e no WRR foram as que se mostraram aptas a oferecer suporte à classe B, garantindo pequenos atrasos para ambas as sub-classes. A arquitetura baseada em WRR foi a que permitiu o atendimento simultâneo a uma menor quantidade de fluxos para um mesmo nível de serviço.

Foi realizada também a análise da utilização de um mecanismo de controle de admissão de chamadas baseado em medidas (MBAC) nos ambientes móveis. O mecanismo foi sintonizado e avaliado independentemente da arquitetura utilizada (SPP, PQ ou WRR), demonstrando resultados semelhantes para diferentes níveis de mobilidade. A integração do MBAC com os mecanismos de suporte às classes demonstrou o correto funcionamento em termos de limitação da degradação dos serviços oferecidos. A integração do MBAC com o WRR foi a que se mostrou mais sensível à correta configuração do conjunto de mecanismos.

Como continuação desse trabalho, podem ser investigados novos mecanismos que possam dar suporte a classes de serviço voltadas para a mobilidade. Podem ser estudados mecanismos que implementem as outras classes propostas no trabalho. A avaliação da integração das diferentes classes propostas deve também ser realizada. Políticas que combinem propriedades de prioridade tanto espacial como temporal oferecem diferenciação de atraso e perdas [91], podendo ser empregadas para a criação de mecanismos de suporte às classes. Cenários de simulação que modelem de maneira mais realista as redes móveis também podem ser estudados.

Referências Bibliográficas

[1] Guardini, I., D'Urso, P., and Fasano, P. The Role of Internet Technology in Future Mobile Data Systems. *IEEE Communications Magazine 38*, 11 (2000), 68–72.

[2] Barden, R., Clark, D. D., and Shenker, S. Integrated Services in the Internet Archictecture: An Overview. *Internet RFC 1633* (junho de 1994).

[3] Blake, S., Black, D. L., Carlson, M., Davies, E., Wang, Z., and Weiss, W. An Architecture for Differentiated Services. *Internet RFC 2475* (dezembro de 1998).

[4] Nichols, K., Blake, S., Baker, F., and Black, D. L. Definition of the Differentiated Services Field (DS Field) in the IPv4 and IPv6 Headers. *Internet RFC 2474* (dezembro de 1998).

[5] The Internet Engineering Task Force web site,. URL: http://www.ietf.org/.

[6] Singh, S. Quality of Service Guarantees in Mobile Computing. *Journal of Computer Communications 19*, 4 (1996), 359–371.

[7] Seal, K., and Singh, S. Loss profiles: A Quality of Service Measure in Mobile Computing. *Journal of Wireless Networks 2*, 1 (1996), 45–61.

[8] Cidon, I., Guérin, R., and Khamisy, A. On protective buffer policies. In *Proceedings of INFOCOM '93* (1993), pp. 1051–1058.

[9] Keshav, S. *An Engineering Approach to Computer Networking.* Addison-Wesley, 1997.

[10] Katz, R. H. Adaptation and Mobility in Wireless Information Systems. *IEEE Personal Communications* (1994), 6–17.

[11] Chalmers, D., and Sloman, M. A Survey of Quality of Service in Mobile Computing Environments. *IEEE Communications Surveys & Tutorials 2*, 2 (1999).

[12] Cordeiro, C. M., and Agrawal, D. P. Mobile Ad hoc Networking. *Livro Texto dos Minicursos - SBRC'2002* (Maio 2002).

[13] Tanenbaum, A. S. *Computer Networks*, 3rd. ed. Prentice Hall, New Jersey, 1996.

[14] Dias, K. L., and Sadok, D. F. H. Internet Móvel: Tecnologias, Aplicações e QoS. *Livro Texto dos Minicursos - SBRC'2001* (2001).

[15] Eckhardt, D., and Steenkiste, P. Measurement and Analysis of the Error Characteritics of an In-Building Wireless Network. In *Proceedings of ACM SIGCOMM Symposium on Communication Architectures and Protocols* (EUA, agosto de 1996).

[16] Forman, G. H., and Zahorjan, J. The Challenges of Mobile Computing. Relatório técnico, CSE - University of Washington, março de 1994. `URL: ftp://ftp.cs.washington.edu`.

[17] Nandagopal, T., Kim, T., Sinha, P., and Bharghavan, V. Service Differentiation Through End-to-end Rate Control in Wireless Packet Networks. In *Proceedings of IEEE Mobile Multimedia Conference* (1999).

[18] Postel, J. Transmission Control Protocol. *Internet RFC 793* (setembro de 1981).

[19] Schiller, J. *Mobile Communications*. Addison-Wesley Publishing Company, 2000.

[20] Ramjee, R., Porta, T. F. L., Thuel, S., Varadhan, K., and Wang, S. Y. HAWAII: A Domain-based Approach for Supporting Mobility in Wide-

Area Wireless Networks. In *Seventh International Conference on Network Protocols* (Toronto, Canada, 1999), pp. 283–292.

[21] Perkins, C. E. Mobile IP. *IEEE Communications Magazine 35* (maio de 1997), 84–99.

[22] Campbell, A., Gomez, J., Kim, S., Valko, A., Wan, C., and Turanyi, Z. Design, Implementation, and Evaluation of Cellular IP, agosto de 2000.

[23] Satyanarayanan, M. Fundamental Challenges in Mobile Computing. In *Symposium on Principles of Distributed Computing* (1996), pp. 1–7.

[24] Fieger, A., Zitterbart, M., Keller, R., and Diederich, J. Towards QoS-support in the Presence of Handover. In *First Workshop on IP Quality of Service for Wireless and Mobile Networking, IQWiM'99* (Germany, abril de 1999).

[25] Brewer, E. A., and Katz, R. H. A Network Architecture for Heterogeneous Mobile Computing. *IEEE Personal Communications* (outubro de 1998).

[26] Kalden, R., Meirik, I., and Meyer, M. Wireless Internet Access Based on GPRS. *IEEE Personal Communications 7*, 2 (abril de 2000), 8–18.

[27] Manner, J., López, A., Mihailovic, A., Velayos, H., Hepworth, E., and Khouaja, Y. Evaluation of Mobility and QoS Interaction. *Computer Networks 38*, 2 (fevereiro de 2002), 137–163.

[28] Kempf, J., McCann, P., and Roberts, P. IP Mobility and the CDMA Radio Access Network: Applicability Statement for Soft Handoff. *Internet Draft (work in progress)* (setembro de 2001).

[29] Comer, D. E. *Internetworking with TCP/IP*, vol. 1. Prentice Hall, New Jersey, 1995.

[30] Liao, W., and Liu, J.-C. VoIP Mobility in IP/cellular Network Internetworking. *IEEE Communications Magazine 38*, 4 (abril de 2000), 70–75.

[31] Ramjee, R., Porta, T. F. L., Salgarelli, L., Thuel, S., and Varadhan, K. HAWAII: A Domain-based Approach for Supporting Mobility in Wide-Area Wireless Networks. *IEEE Personal Communications* (agosto de 2000), 34–41.

[32] Chen, J.-C., McAuley, A., Caro, A., Baba, S., Ohba, Y., and Ramanathan, P. QoS Architecture Based on Differentieted Services for Next Generation Wireless IP Networks. *Internet Draft (work in progress)* (julho de 2000).

[33] Dixit, S., Guo, Y., and Antoniou, Z. Resource management and Quality so Service in third-generation wireless networks. *IEEE Communications Magazine 39*, 2 (fevereiro de 2001), 125–133.

[34] Huber, J. F., Weiler, D., and Brand, H. UMTS, the Mobile Multimedia Vision for IMT-2000: A Focus on Standadization. *IEEE Communications Magazine 38*, 9 (setembro de 2000), 129–136.

[35] Bos, L., and Leroy, S. Toward an All-IP-Based UMTS System Architecture. *IEEE Network Magazine 15*, 1 (janeiro de 2001), 36–45.

[36] 3GPP web site,. URL: http://www.3gpp.org/.

[37] Priggouris, G., Hadjiefthymiades, S., and Merakos, L. Supporting IP QoS in the general packet radio service. *IEEE Network Magazine* (setembro de 2000), 8–17.

[38] Sarikaya, B. Packet Mode in Wireless Networks: Overview of Transition to Third Generation. *IEEE Communications Magazine 38*, 9 (setembro de 2000), 164–172.

[39] Partain, D., Karagiannis, G., Wallentin, P., and Westberg, L. Resource reservation issues in cellular access networks. *Internet Draft (work in progress)* (abril de 2001).

[40] Karagiannis, G., Partain, D., Rexhepi, V., and Westberg, L. QoS Signalling Requirements for Wireless Networks. *Internet Draft (work in progress)* (novembro de 2001).

[41] Koodli, R., and Puuskari, M. Supporting packet-data QoS in next-generation cellular networks. *IEEE Communications Magazine 39*, 2 (fevereiro de 2001), 180–188.

[42] Technical Specification Group Services and System Aspects. QoS Concept and Architecture - Technical Specification 3GPP TS 23.107 V5.5.0 (2002-06). *3rd Generation Partnership Project* (2002).

[43] Kurose, J. F., and Ross, K. W. *Computer Networking: A Top-Down Approach Featuring the Internet.* Addison-Wesley Publishing Company, 2001.

[44] Lu, S., and Bharghavan, V. Adaptive Resource Management Algorithms for Indoor Mobile Computing. In *Proceedings of ACM SIGCOMM Symposium on Communication Architectures and Protocols* (EUA, agosto de 1996), pp. 231–242.

[45] Hassan, M., and Nayandoro, A. Internet Telephony: Services, Techincal Challenges and Products. *IEEE Communications Magazine 38*, 4 (abril de 2000), 96–103.

[46] Diederich, J., and Zitterbart, M. An Expedited Forwarding with dropping PHB. *Internet Draft (work in progress)* (outubro de 1999).

[47] Das, S. K., Sen, S. K., Agrawal, P., and Basu, K. Modeling QoS Degradation in Multimedia Wireless Networks. In *Proceedings of IEEE International Conference on Personal Wireless Communications* (1997).

[48] Shenker, S., Partridge, C., and Guerin, R. Specification of guaranteed quality of service. *Internet RFC 2212* (setembro de 1997).

[49] Wroclawski, J. Specification of the controlled-load network element service. *Internet RFC 2211* (setembro de 1997).

[50] Mathy, L., Edwards, C., and Hutchison, D. The Internet: A Global Telecommunications Solution? *IEEE Network Magazine 14*, 4 (julho de 2000), 46–57.

[51] Zhang, L., Deering, S. E., Estrin, D., Shenker, S., and Zappala, D. RSVP: A new resource reservation protocol. *IEEE Network 7*, 5 (setembro de 1993), 8–18.

[52] Braden, R., Zhang, L., Berson, S., Herzog, S., and Jamin, S. Resource reservation protocol (RSVP) – version 1 functional specification. *Internet RFC 2205* (setembro de 1997).

[53] Mankin, A., Baker, F., Braden, R., Bradner, S., O'Dell, M., Romanow, A., Weinrib, A., and Zhang, L. Resource reservation protocol (RSVP) – version 1 applicability statement. *Internet RFC 2208* (setembro de 1997).

[54] Wroclawski, J. The use of RSVP with IETF integrated services. *Internet RFC 2210* (setembro de 1997).

[55] Bernet, Y., Ford, P., Yavatkar, R., Baker, F., Zhang, L., Speer, M., Braden, R., Wroclawski, B. D. J., and Felstaine, E. A Framework for Integrated Services Operation over Diffserv Networks. *Internet RFC 2998* (novembro de 2000).

[56] Nichols, K., and Carpenter, B. Definition of Differentiated Services Per Domain Behaviors and Rules for their Specification. *Internet RFC 3086* (abril de 2001).

[57] Davie, B., Charny, A., Bennett, J., Benson, K., Boudec, J. L., Courtney, W., Davari, S., Firoiu, V., and Stiliadis, D. An Expedited Forwarding PHB (Per-Hop Behavior). *Internet RFC 3246* (março de 2002).

[58] Charny, A., Baker, F., Davie, B., Bennett, J. C. R., Benson, K., Boudec, J. Y. L., Chiu, A., Courtney, W., Davari, S., Firoiu, V., Kalmanek, C., Ramakrishnan, K. K., and Stiliadis, D. Supplemental

Information for the New Definition of the EF PHB (Expedited Forwarding Per-Hop Behavior). *Internet RFC 3247* (março de 2002).

[59] HEINANEN, J., BAKER, F., WEISS, W., AND WROCLAWSKI, J. Assured Forwarding PHB group. *Internet RFC 2597* (junho de 1999).

[60] FLOYD, S., AND JACOBSON, V. Random Early Detection Gateways for Congestion Avoidance. *IEEE/ACM Transactions on Networking 1*, 4 (agosto de 1993), 397–413.

[61] DE MEER, H., FEHER, G., BLEFARI-MELAZZI, N., KARAGIANNIS, G., PARTAIN, D., AND WESTBERG, L. Analysis of Existing QoS Solutions. *Internet Draft (work in progress)* (novembro de 2001).

[62] DE MEER, H., AND O'HANLON, P. Segmented Adaptation of Traffic Aggregates. *Lecture Notes in Computer Science 2092* (2001).

[63] HUSTON, G. Next Steps for the IP QoS Architecture. *Internet RFC 2990* (novembro de 2000).

[64] BAKER, F., ITURRALDE, C., FAUCHEUR, F. L., AND DAVIE, B. Aggregation of RSVP for IPv4 and IPv6 Reservations. *Internet RFC 3175* (setembro de 2001).

[65] WESTBERG, L., JACOBSSON, M., KARAGIANNIS, G., OOSTHOEK, S., PARTAIN, D., REXHEPI, V., SZABO, R., AND WALLENTIN, P. Resource Management in Diffserv (RMD) Frameworkk. *Internet Draft (work in progress)* (fevereiro de 2002).

[66] EHRENSBERGER, J. Resource Demand of Aggregated Resource Reservations. In *Proceedings of the 1st IEEE European Conference on Universal Multiservice Networks (ECUMN'2000)* (Colmar, France, outubro de 2000).

[67] HEIJENK, G., KARAGIANNIS, G., REXHEPI, V., AND WESTBERG, L. DiffServ Resource Management in IP-based Radio Access Networkss. In *Proceedings of 4th International Symposium on Wireless Personal Multimedia Communications (WPMC'01)* (Aalborg, Denmark, setembro de 2001).

[68] Jacobsson, M., Karagiannis, G., de Kogel, M., Oosthoek, S., Partain, D., Rexhepi, V., and Wallentin, P. Resource Management in Diffserv On DemAnd (RODA) PHR. *Internet Draft (work in progress)* (fevereiro de 2002).

[69] Kim, J., and Jamalipour, A. Traffic Management and QoS Provisioning in Future Wireless IP Networks. *IEEE Personal Communications* (outubro de 2001), 46–55.

[70] Szabó, R., Henk, T., Rexhepi, V., and Karagiannis, G. Resource Management in Differentiated Services (RMD) IP Networks. In *Proceeding of the International Conference on Emerging Telecommunications Technologies and Applications* (Kosice, Slovak Republic, 2001).

[71] Liao, R. R. F., and Campbell, A. Dynamic Edge Provisioning for Core IP Networks. In *Proceedings of the 8th International Workshop on Quality of Service - IWQoS 2000* (Pittsburg, USA, junho de 2000).

[72] Liao, R. R. F., and Campbell, A. A Dynamic Provisioning Model for Differentiated Service. Relatório técnico, Columbia Univerity, 2000.

[73] Wroclawski, J., and Charny, A. Integrated Service Mappings for Differentiated Services Networks. *Internet Draft (work in progress)* (2001).

[74] Breslau, L., Knightly, E. W., Shenker, S., Stoica, I., and Zhang, H. Endpoint admission control: Architectural issues and performance. In *Proceedings of ACM SIGCOMM 2000* (Stockholm, Sweden, agosto de 2000), pp. 57–69.

[75] Chuah, C.-N., Subramanian, L., Katz, R. H., and Joseph, A. D. QoS Provisioning Using a Clearing House Architecture. In *Proceedings of the 8th International Workshop on Quality of Service - IWQoS 2000* (Pittsburg, USA, junho de 2000).

[76] Terzis, A., Wang, L., and Zhang, L. A Two-Tier Resource Management Model for the Internet. In *Proceedings of Global Internet* (Rio de Janeiro, Brazil, dezembro de 1999).

[77] de Vasconcellos, S. V., and de Rezende, J. F. Mobilidade num Ambiente de Serviços Diferenciados. In *Anais do XIX Simpósio Brasileiro de Redes de Computadores, SBRC'2001* (Florianópolis, SC, maio de 2001), pp. 242–257.

[78] Knightly, E. W., and Shroff, N. B. Admission Control for Satistical QoS: Theory and Practice. *IEEE Network Magazine* (março de 1999), 20–29.

[79] Srikant, R. Control of Communication Networks. *Perspectives in Control Engineering: Technologies, Applications, New Directions* (2000), 462–488.

[80] Floyd, S. Comments on measurement-based admission control for controlled-load services. Relatório técnico, julho de 1996. Submitted for CCR.

[81] el Allali, H., and Heijenk, G. Resource Management in IP-based Radio Access Networks. In *CTIT Workshop on Mobile Communications* (fevereiro de 2001).

[82] Breslau, L., Jamin, S., and Shenker, S. Comments on the Performance of Measurement-Based Admission Control Algorithms. In *Proceedings of INFOCOM'00* (março de 2000), pp. 1233–1242.

[83] Jamin, S., Shenker, S. J., and Danzig, P. B. Comparison of Measurement-based Admission Control Algorithms for Controlled-Load Service. In *Proceedings of the Conference on Computer Communications (INFOCOM'97)* (abril de 1997).

[84] Jamin, S., Danzig, P. B., Shenker, S. J., and Zhang, L. A Measurement-based Admission Control Algorithm for Integrated Services PAcket Networks (Extended Version). *IEEE/ACM Transactions on Networking* (fevereiro de 1997).

[85] Fall, K., and Varadhan, K. Network simulator - NS (version 2), the ns manual. Relatório técnico, The VINT Project, 2002. URL: http://www.isi.edu/nsnam/ns/ .

[86] Ziviani, A., de Rezende, J. F., and Duarte, O. C. M. B. Tráfego de Voz em um Ambiente de Diferenciação de Serviços na Internet. In *Anais do XVII Simpósio Brasileiro de Redes de Computadores, SBRC'99* (Salvador, BA, maio 1999), pp. 354–367.

[87] de Rezende, J. F. Avaliação do Serviço Assegurado para a Diferenciação de Serviços na Internet. In *Anais do XVII Simpósio Brasileiro de Redes de Computadores, SBRC'99* (Salvador, BA, maio 1999), pp. 339–353.

[88] Jain, R. *The Art of Computer Systems Performance Analysis.* John Wiley and Sons Inc., 1991.

[89] Brandy, P. A Technique for Investigating On/Off Patterns of Speech. *Bell Labs Tech Journal 44*, 1 (janeiro de 1965), 1–22.

[90] de Vasconcellos, S. V., and de Rezende, J. F. Using Differentiated Services in 3G Cellular Networks. In *Proceedings of the 3rd. IEEE International Conference on Mobile and Wireless Communication Networks, MWCN'2001* (Recife, PE, agosto de 2001).

[91] Gravey, A., and Hébuterne, H. Mixing Time and Loss Priorities in a Single Server Queue. In *Proceedings of ITC-13* (Denmark, junho de 1990), pp. 147–152.