

MODELAGEM E ANÁLISE DE UM PROTOCOLO DE ACESSO
ALTERNATIVO PARA O PADRÃO IEEE 802.16 DE REDES
METROPOLITANAS SEM FIO

Paulo Ditarso Maciel Júnior

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 DE SISTEMAS E COMPUTAÇÃO.

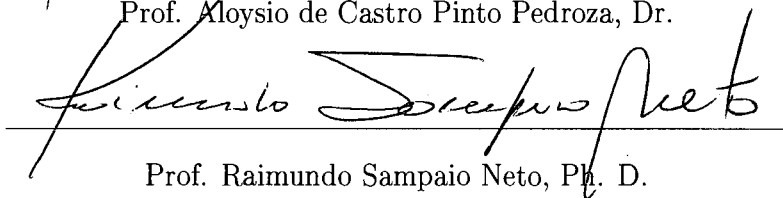
Aprovada por:



Prof. Luís Felipe Magalhães de Moraes, Ph. D.



Prof. Aloysio de Castro Pinto Pedroza, Dr.



Prof. Raimundo Sampaio Neto, Ph. D.

RIO DE JANEIRO, RJ - BRASIL

ABRIL DE 2005

MACIEL JÚNIOR, PAULO DITARSO

Modelagem e Análise de um Protocolo de Acesso Alternativo para o Padrão IEEE 802.16 de Redes Metropolitanas sem Fio [Rio de Janeiro] 2005

XIV, 86 p. 29,7 cm (COPPE/UFRJ, M.Sc., Engenharia de Sistemas e Computação, 2005)

Tese - Universidade Federal do Rio de Janeiro, COPPE

1. Redes Metropolitanas Sem Fio
2. Protocolos de Acesso ao Meio
3. Qualidade de Serviço
4. Avaliação de Desempenho

I. COPPE/UFRJ II. Título (série)

Dedicatória

Dedico esse trabalho a minha sobrinha Mirella Maciel. Espero que um dia você possa perdoar todos os dias em que estive ausente e entender o quanto eu te amo, mesmo distante.

Agradecimentos

Primeiramente, gostaria de agradecer a Deus pelo dom da vida e pela graça de poder realizar este trabalho.

Aos meus pais Paulo e Ilma, aos meus irmãos Sérgio e Suênia, a minha linda sobrinha Mirella, bem como, toda a minha família (em especial a querida tia Heremita, pelo apoio incondicional ao meu estudo), por todo amor e carinho, não só durante a realização deste trabalho, como na vida inteira. Agradecimento mais que especial a minha mãe, pelo apoio, amor e preocupação, sempre. Sem você eu não teria conseguido.

Agradeço ao meu orientador, Prof. Luís Felipe, pelos grandes ensinamentos e pelo total apoio desde o início do meu trabalho e aos demais integrantes da banca, os Professores Aloysio Pedroza e Raimundo Sampaio, pela valiosa ajuda nesta fase final.

Agradeço a família Costa (Marcone, Cristina, Marcone Filho, Carlota, Mônica, Waldney e Miguel), pelo apoio no início da minha jornada e pela grande amizade que foi selada.

Agradeço a Maria Otília, Carlos Walney e Carlinhos, pela acolhida e pelo carinho com que me tratam.

Agradeço a todos os amigos que conheci durante este período, em particular: Bruno, Pinaffi, Denilson, Victor, Luciano, Daniel, Caio, Mendes, Villela, Eduardo, Airon, Júlio, Cláudia, Marcos, Diogo, Michelini, Flávia, Demétrio, Júnior, Rafael, Guto, Bernardo, Michael, prof. Manuel, Izabela e todos os outros, que por ventura eu tenha esquecido. Agradecimento especial ao amigo Beto, não só pelos ensinamentos transmitidos, como também pelo seu companheirismo e amizade.

Agradeço a família Tomimura (Diana, Patrícia, Nazira) por todo amor e carinho. Em especial, gostaria de agradecer a Diana Tomimura, não só pela valiosa

contribuição ao meu trabalho, como também, pela companhia nos melhores e piores momentos. Obrigado pelo seu apoio nas horas difíceis, pela sua alegria nos momentos de tristeza, pelo seu carinho nos momentos frígidos, pela sua ternura nos momentos complicados, e, acima de tudo, pelo seu amor. Espero um dia poder retribuir toda felicidade que você me proporcionou.

À Fundação Carlos Chagas Filho de Amparo à Pesquisa do Estado do Rio de Janeiro (FAPERJ), pelo financiamento da pesquisa e ao Programa de Engenharia de Sistemas e Computação (PESC/COPPE/UFRJ), pelo apoio operacional.

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

MODELAGEM E ANÁLISE DE UM PROTOCOLO DE ACESSO
ALTERNATIVO PARA O PADRÃO IEEE 802.16 DE REDES
METROPOLITANAS SEM FIO

Paulo Ditarso Maciel Júnior

Abril/2005

Orientador: Luís Felipe Magalhães de Moraes

Programa: Engenharia de Sistemas e Computação

Os Sistemas BWA (Broadband Wireless Access) surgiram como uma solução para o acesso à banda-larga através de meios sem fio. Estes sistemas foram também desenvolvidos para transmitir dados e serviços multimídia com diferentes requisitos de qualidade de serviço (QoS). O IEEE 802.16 especifica a camada PHY e MAC para sistemas BWA. Porém, o padrão prevê apenas o suporte a QoS e não define como escalonar os diferentes tipos de tráfego. Neste trabalho é proposto um novo protocolo MAC para BWA que incorpora um mecanismo de escalonamento de tráfego com prioridades baseadas em mensagens e/ou em estações. Além disso, um modelo analítico para o tempo de espera das mensagens é apresentado e, através deste modelo, alguns resultados numéricos são obtidos.

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

MODELING AND ANALYSIS OF AN ALTERNATIVE ACCESS PROTOCOL
FOR THE IEEE 802.16 STANDARD OF WIRELESS METROPOLITAN AREA
NETWORKS

Paulo Ditarso Maciel Júnior

April/2005

Advisor: Luís Felipe Magalhães de Moraes

Department: Systems Engineering and Computer Science

The Broadband Wireless Access Systems (BWA) appeared as a solution for broadband access through wireless network. This systems was developed to transmit data and multimedia services with distincts quality-of-service (QoS) requirements. IEEE 802.16 Standard specifies the PHY and MAC layers for BWA systems. However, the standard defines only QoS suport but not how to schedule different types of traffic. This work propose a new MAC protocol for BWA systems that incorporates a traffic scheduling mecanism based on messages and/or stations priorities. Moreover, an analitic model of the messages waiting time and some numeric results are presented.

Lista de Acrônimos

ATM	:	<i>Asynchronous Transfer Mode;</i>
BS	:	<i>Base Station;</i>
DL	:	<i>Downlink;</i>
DSL	:	<i>Digital Subscriber Line;</i>
ETSI	:	<i>European Telecommunications Standards Institute;</i>
FDD	:	<i>Frequency Division Duplex;</i>
FIFO	:	<i>First In, First Out;</i>
FCFS	:	<i>First Come, First Served;</i>
HOL	:	<i>Head of the Line;</i>
IEEE	:	<i>Institute of Electrical and Electronic Engineers;</i>
MAC	:	<i>Media Access Control;</i>
PHY	:	<i>Physical Layer;</i>
QoS	:	<i>Quality of Service;</i>
SS	:	<i>Subscriber Station;</i>
TDD	:	<i>Time-Division Duplex;</i>
TDM	:	<i>Time-Division Multiplexed;</i>
TDMA	:	<i>Time-Division Multiple Access;</i>
UL	:	<i>Uplink;</i>
WLAN	:	<i>Wireless Local Area Network;</i>
WMAN	:	<i>Wireless Metropolitan Area Network;</i>
WPAN	:	<i>Wireless Personal Area Network;</i>

Conteúdo

Resumo	vi
Abstract	vii
Lista de Acrônimos	viii
Lista de Figuras	xii
Lista de Tabelas	xiv
1 Introdução	1
1.1 Redes sem Fio	2
1.2 Protocolos de Múltiplo Acesso	4
1.3 Objetivo do Trabalho	8
1.4 Contribuições	9
1.5 Estrutura da Tese	9
2 Qualidade de Serviço	11
2.1 Visão Geral	12
2.2 Tipos de Tráfego	14

<i>CONTEÚDO</i>	x
2.3 Escalonamento de Pacotes	16
2.3.1 <i>First-In-First-Out</i> (FIFO)	16
2.3.2 Fila de Prioridades	16
2.3.3 <i>Weighted Fair Queuing</i> (WFQ)	17
2.4 Controle de Admissão	18
2.5 Policiamento de Tráfego	19
3 Padrão IEEE 802.16	20
3.1 Visão Geral	21
3.2 Camada PHY e MAC	22
3.3 Arquitetura de QoS	25
3.4 Wi-Fi <i>versus</i> WiMAX	27
4 Trabalhos Relacionados	28
4.1 Protocolos de Reserva	29
4.1.1 Notação Matemática	30
4.1.2 MLMA	31
4.1.3 RPAC	34
4.2 Qualidade de Serviço em 802.16	38
5 Protocolo Proposto	41
5.1 Descrição	42
5.2 Comentários	46
6 Modelagem Analítica	48

<i>CONTEÚDO</i>	xi
6.1 Modelagem da Versão I	49
6.2 Modelagem da Versão II	55
7 Resultados Obtidos	58
7.1 Cenários de Avaliação	59
7.2 Validação do Modelo	70
8 Conclusões e Trabalhos Futuros	73
8.1 Conclusões	74
8.2 Trabalhos Futuros	75
Bibliografia	76
A Sistema M/G/1	82

Lista de Figuras

1.1	Padronizações da tecnologia de redes sem fio.	3
1.2	Classificação dos protocolos MAC.	6
2.1	Abstração de um fila FIFO.	16
2.2	Modelo de fila com prioridades.	17
2.3	<i>Weighted Fair Queuing</i> (WFQ).	17
3.1	Variedade de estações clientes comunicando-se com uma estação base.	21
3.2	Arquitetura básica do sistema BWA.	22
3.3	Estrutura do Quadro TDD	23
3.4	Estrutura do quadro MAC no esquema TDD.	24
3.5	Estrutura de alocação do 802.16.	25
3.6	Arquitetura de QoS do IEEE 802.16.	27
4.1	Esquema de acesso.	31
4.2	Comparação das disciplinas de prioridade: fixa, cíclica e complementar.	33
4.3	Comportamento do RPAC I.	35
4.4	Comportamento do RPAC II.	36
4.5	Tempo médio de espera por estação: RPAC I (a) e RPAC II (b). . . .	37

5.1	Estrutura do quadro MAC para o protocolo proposto.	44
5.2	Ciclos consecutivos de transmissão.	44
5.3	Comportamento da Versão I do protocolo proposto.	46
5.4	Comportamento da Versão II do protocolo proposto.	46
7.1	Tempo médio de espera na fila para cada classe de prioridade versus tráfego oferecido no Cenário I: Versão I (a) e Versão II (b).	61
7.2	Tempo médio de espera na fila para cada classe de prioridade versus tráfego oferecido no Cenário II: Versão I (a) e Versão II (b).	62
7.3	Tempo médio de espera na fila para diferentes valores de ρ em cada estação no Cenário I: Versão I (a) e Versão II (b).	63
7.4	Tempo médio de espera na fila para diferentes valores de ρ em cada estação no Cenário II: Versão I (a) e Versão II (b).	64
7.5	Tempo médio de espera na fila para cada classe de prioridade versus tráfego oferecido, com prioridade cíclica: Cenário I (a) e Cenário II (b).	66
7.6	Tempo médio de espera na fila para cada classe de prioridade versus tráfego oferecido, com prioridade complementar: Cenário I (a) e Cenário II (b).	67
7.7	Tempo médio de espera na fila para diferentes valores de ρ em cada estação, com prioridade cíclica: Cenário I (a) e Cenário II (b).	68
7.8	Tempo médio de espera na fila para diferentes valores de ρ em cada estação, com prioridade complementar: Cenário I (a) e Cenário II (b).	69
7.9	Abstração de um sistema M/G/1 com férias e prioridades.	70
7.10	Comparação entre os resultados obtidos analiticamente e por simulação.	72
A.1	Sistema M/G/1 com férias.	84

Lista de Tabelas

1.1	Comparação entre os protocolos de múltiplo acesso e o custo associado.	8
2.1	Requisitos de algumas aplicações de redes (valores aproximados). . . .	15
7.1	Cenários de tráfego utilizados na modelagem analítica.	59

Capítulo 1

Introdução

ESTE capítulo apresenta os conceitos básicos inerentes às comunicações sem fio. Neste tipo de tecnologia, a grande dificuldade é alocar de maneira eficiente o canal de comunicação entre as estações. Vários métodos de acesso têm sido propostos na literatura com o intuito de melhorar o desempenho das transmissões de dados via dispositivos sem fio. Isto torna a computação móvel cada vez mais presente no nosso cotidiano. A primeira seção apresenta uma visão geral de redes sem fio através das padronizações existentes para esta tecnologia. Na seção seguinte, a vantagem do compartilhamento de recursos é ilustrada através de um exemplo, contextualizando a teoria de filas dentro dos sistemas de computação. Além disso, apresenta uma classificação dos protocolos de múltiplo acesso, comparando-os em relação ao custo envolvido na alocação do canal. Na próxima seção, os objetivos do trabalho são expostos. E, por fim, a estrutura do trabalho está descrita na última seção.

1.1 Redes sem Fio

As redes sem fio formam atualmente uma grande vertente tecnológica, justificada pela busca de praticidade e acessibilidade aos meios de comunicação. O suporte a redes sem fio e a implantação de “*hot spots*” (pontos de acesso à rede sem fio, através dos quais pode-se acessar a Internet) nos principais grandes centros, também são uma grande tendência e vêm apresentando um crescimento extremamente rápido. A rápida proliferação dos dispositivos de computação móveis (como *laptops*, *handhelds* e PDAs) conduziu à uma mudança revolucionária na computação mundial nos últimos anos. A era do computador pessoal (um computador por pessoa) está perdendo espaço para a era da “computação presente”, na qual usuários utilizam, ao mesmo tempo, vários aparelhos eletrônicos através dos quais podem acessar todas as informações necessárias a qualquer hora e em qualquer lugar. A natureza destes aparelhos faz da comunicação através de redes sem fio a solução mais simples para interconectá-los.

Um exemplo disto está no crescimento da utilização da tecnologia sem fio tanto no ambiente de rede local (*Wireless Local Area Network* - WLAN), como para redes metropolitanas (*Wireless Metropolitan Area Network* - WMAN). Além de suportar a conectividade sem fio de estações fixas, portáteis e móveis, dentro de uma determinada área, uma rede sem fio pode oferecer conexão aos serviços oferecidos na Internet. É previsível que em um futuro não muito distante, este tipo de tecnologia será amplamente utilizada como meio de acesso à grande rede.

Uma WLAN tem o mesmo alcance de comunicação (de 100 a 500 metros) e deve satisfazer os mesmos requisitos de uma LAN, incluindo alta capacidade, completa conectividade entre as estações e a capacidade de *broadcast*. Para isto, WLANs devem ser projetadas para cobrir algumas questões específicas de ambientes sem fio, tais como consumo de energia, mobilidade, segurança e limitações na capacidade do canal [1]. Atualmente existem dois padrões bem definidos para redes locais sem fio: o padrão IEEE 802.11 [2] e o padrão ETSI HIPERLAN [3]. O padrão IEEE 802.11 é o mais amplamente difundido através da aliança internacional de fabricantes WECA (*Wireless Ethernet Compatibility Alliance*), que possui 183 companhias associadas

por todo o mundo e mais de 2.000 produtos certificados Wi-Fi (*Wireless Fidelity*) [4].

A crescente demanda por acesso à Internet com alta-velocidade (banda-larga) e pelos serviços de multimídia para clientes residenciais e corporativos, proporcionou o rápido desenvolvimento do acesso sem fio para WMAN. O chamado *Broadband Wireless Access System* (BWA) surgiu como a “última milha” para acesso à banda-larga e apresenta várias vantagens em relação aos sistemas de cabo e DSL (*Digital Subscriber Line*), como por exemplo: rápida implantação, alta escalabilidade, baixo custo de atualização e manutenção, etc. Algumas padronizações para o BWA, como o padrão IEEE 802.16 [5] e o padrão ETSI HIPERMAN [6], surgiram com o intuito de prover um sistema de acesso sem fio de alta velocidade e de alto desempenho, com diferenciação de serviços para tipos de tráfego com diferentes requisitos de qualidade de serviço (*Quality of Service - QoS*). Da mesma forma que o 802.11, o padrão IEEE 802.16 também faz parte de uma aliança internacional denominada WiMAX [7], garantindo a interoperabilidade entre os dispositivos de diferentes fabricantes.

Ainda no escopo da comunicação sem fio, existem as redes pessoais (*Wireless Personal Area Network - WPAN*), abrangendo o ambiente que cerca o usuário, em um alcance de aproximadamente 10 metros. Os aparelhos celulares e PDAs que utilizam o protocolo *bluetooth* [8] são exemplos desta tecnologia. A Figura 1.1 ilustra algumas padronizações de comunicação sem fio, representando a hierarquia de cobertura de serviço.

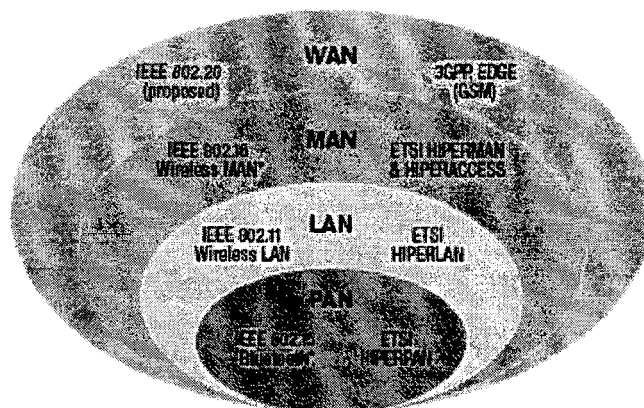


Figura 1.1: Padronizações da tecnologia de redes sem fio.

Um dos grandes problemas que surgem nas redes de comunicações sem fio é como encontrar uma forma econômica e eficiente de compartilhar o recurso mais caro e escasso de uma rede de telecomunicações, o meio de transmissão. O problema neste caso é como controlar o acesso a este canal compartilhado de uma forma que a faixa de transmissão seja dividida eficientemente entre os usuários. A solução mais adequada depende das características do ambiente em questão e dos requisitos que devem ser atendidos. Portanto, uma análise mais detalhada destes sistemas se torna de extrema importância e de grande utilidade, como será visto na próxima seção.

1.2 Protocolos de Múltiplo Acesso

A análise e a aplicação da **teoria de sistemas de filas** [9] podem ser utilizadas em várias áreas, inclusive no campo de sistemas de computação. Um grande problema que recai em análise de fila acontece quando clientes (usuários) competem pelo acesso a um recurso limitado. Este é o problema clássico do compartilhamento e alocação de recursos, que podem ser de vários tipos como: a capacidade de processamento de uma CPU, a utilização de uma memória compartilhada, a capacidade de armazenamento em disco e, no caso das redes sem fio, o canal de comunicação.

Várias questões envolvendo redes de computadores tratam da alocação eficiente do canal de comunicação entre demandas competitivas. A vantagem do compartilhamento de recursos pode ser ilustrada com o seguinte exemplo [10]:

“Imagine dois usuários necessitando utilizar um canal de comunicação. A solução clássica para satisfazê-los é prover um canal dedicado entre eles o tempo que for necessário e cobrá-los pelo uso completo do meio. No entanto, melhor do que disponibilizar vários canais dedicados entre os usuários, é prover um único canal de alta velocidade para ser compartilhado por um grande número de usuários.

Esta vantagem vem da **lei dos grandes números** [9] a qual declara que, com uma alta probabilidade, a demanda em qualquer instante será muito próxima a soma média das demandas daquela população de usuários.”

Em sistemas de comunicação existe a grande necessidade de compartilhar recursos de alto custo entre uma coleção de usuários caracterizados pelo tráfego em rajada. Dentro do contexto deste trabalho, os princípios gerais do compartilhamento de recursos fornecem as principais características da comunicação através de redes sem fio (baixo custo, alta eficiência, flexibilidade, etc).

Quando um recurso é compartilhado por vários usuários independentes, existe a necessidade de utilizar protocolos de múltiplo acesso para coordenar o compartilhamento. No escopo deste trabalho, o recurso que deseja-se compartilhar é o canal de comunicação sem fio entre as estações. Em ambientes dinâmicos, tais como nas comunicações sem fio, é necessário encontrar uma forma de compartilhar o canal de maneira adaptativa. Além das questões relacionadas a teoria de filas [9] devido à natureza aleatória das demandas, alocar o canal para um conjunto de demandas geograficamente distribuídas (e possivelmente móveis) é um sério problema e tem um custo associado. Este custo pode aparecer na forma de colisões devido à um fraco (ou nenhum) controle, capacidade de transmissão desperdiçada por causa de um controle muito rígido ou sobrecarga adicional no tráfego do sistema em função de um controle dinâmico [11]. Portanto, existe desperdício devido ao custo de organizar as demandas em algum tipo de fila cooperativa, que permita o acesso eficiente ao recurso disponível.

Classificação

Os diversos protocolos de controle de acesso ao meio (*Media Access Control* - MAC) existentes podem ser classificados de acordo com [12]: suas funcionalidades em relação a natureza estática ou dinâmica do canal, um mecanismo de controle centralizado ou distribuído, e o comportamento adaptativo do algoritmo de controle.

Esta classificação está ilustrada na Figura 1.2¹. Cada classe tem suas próprias vantagens e desvantagens, e não existe nenhum protocolo que se sobressaia aos outros sob todos os aspectos de desempenho.

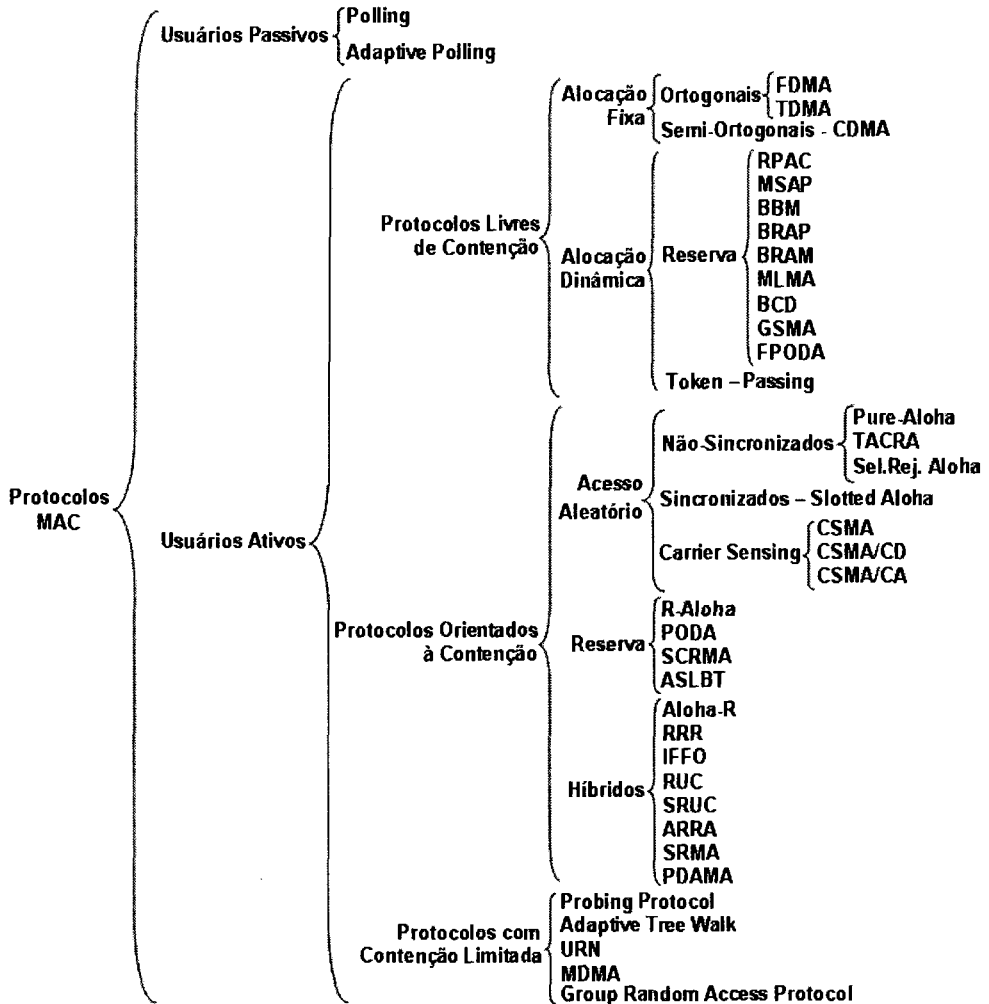


Figura 1.2: Classificação dos protocolos MAC.

De acordo com o esquema de controle do algoritmo de múltiplo acesso, estes protocolos podem ser sub-classificados em três categorias segundo [15]:

- Protocolos de Alocação Fixa
- Protocolos de Acesso Aleatório
- Protocolos de Alocação Dinâmica

¹Para um estudo mais detalhado sobre protocolos de múltiplo acesso, recomenda-se [13, 14, 12].

Os protocolos de alocação fixa caracterizam-se por atribuir uma parte do canal para cada estação, de maneira fixa. O TDMA (*Time-Division Multiple Access*) e o FDMA (*Frequency-Division Multiple Access*) [13] são exemplos deste tipo de protocolo. São extremamente fáceis de implementar, porém, quando uma estação não tem pacotes para enviar durante o período de tempo em que o meio está alocado para ela, o canal ficará ocioso enquanto outros terminais poderiam estar utilizando-o.

Por outro lado, os protocolos de acesso aleatório não possuem um controle rígido para alocação do canal. O ALOHA [16] e o CSMA (*Carrier Sense Multiple Access*) [17] são exemplos desta categoria. Também são relativamente simples de implementar, porém, a possibilidade de ocorrer colisões quando duas ou mais estações tentam transmitir ao mesmo tempo provoca desperdício no canal de transmissão.

Nos protocolos de alocação dinâmica o canal é alocado de acordo com as necessidades das estações. Este controle pode ser implementado de várias maneiras, como por exemplo: na forma de *polling* [18] (onde uma estação espera ser “questionada” se necessita acessar o canal), ou na forma de reservas explícitas, como no RPAC (*Reservation-Priority Access Control*) [19]. Com a utilização destes protocolos não existe a possibilidade de colisões e o canal é alocado sob demanda, evitando períodos de tempo ociosos no canal. Porém, há uma sobrecarga na rede devido aos sinais de controle transmitidos no meio.

O custo da alocação do canal pode ser classificado em três categorias distintas de acordo com o desperdício associado a cada método: canal ocioso sem transmissão de dados, colisão de pacotes transmitidos ao mesmo tempo, ou sobrecarga de informação de controle no tráfego da rede. A Tabela 1.1 apresenta uma comparação entre os protocolos de múltiplo acesso e o custo devido ao seu funcionamento. Além disso, existe a possibilidade de combinar alguns destes protocolos para formar métodos de acesso híbridos [13]. Estes esquemas sofrem uma combinação dos custos relacionados ao desperdício do canal.

Outra característica interessante é que alguns protocolos de múltiplo acesso já incorporam mecanismos de escalonamento de pacotes através de prioridades, possibilitando uma diferenciação no tratamento de classes de tráfego distintas, como será

	Canal ocioso	Colisões	Sobrecarga
Alocação Fixa	Sim	Não	Não
Acesso Aleatório	Não	Sim	Não
Alocação Dinâmica	Não	Não	Sim

Tabela 1.1: Comparação entre os protocolos de múltiplo acesso e o custo associado.

visto mais adiante. Esta diferenciação é fundamental para transmissão de dados e serviços multimídia com diferentes requisitos de qualidade de serviço (QoS).

1.3 Objetivo do Trabalho

Como será visto mais adiante, ao mesmo tempo que o padrão IEEE 802.16 suporta apenas mecanismos para prover QoS, ele não inclui uma solução completa para fornecer garantias as várias aplicações e não define como escalonar eficientemente o tráfego para satisfazer tais requisitos. Dessa forma, é necessário a introdução de algum mecanismo de escalonamento para que haja uma diferenciação dos serviços oferecidos pelo 802.16. Além disso, garantias de QoS podem ser respeitadas através da implantação de um rígido controle de admissão e policiamento de tráfego.

O objetivo desta tese é apresentar um protocolo de acesso ao meio alternativo para redes metropolitanas sem fio (WMAN) que incorpora funções de escalonamento de tráfego com prioridades baseadas em mensagens e/ou em estações. O protocolo proposto pode ser facilmente implementado no padrão IEEE 802.16 [5] devido à similaridade com o funcionamento da camada MAC do 802.16, como será visto no Capítulo 5. Diversos mecanismos de escalonamento têm sido propostos na literatura para o 802.16. Porém, estas propostas abordam apenas algoritmos para o escalonamento do tráfego e poucos trabalhos têm sido apresentados envolvendo alterações no protocolo de acesso ao meio, para que o tráfego escalonado possa ser eficientemente escoado.

Além disso, este trabalho apresenta um modelo analítico para o tempo médio de espera na fila nas duas versões do protocolo proposto, considerando chegada de

mensagens de diferentes tipos de tráfego em cada estação. Para avaliar a solução proposta, os resultados obtidos com o modelo analítico são comparados com resultados obtidos através de simulação.

1.4 Contribuições

Dentre os principais resultados alcançados com a elaboração deste trabalho, as seguintes contribuições podem ser relacionadas:

- A elaboração de um protocolo de acesso ao meio alternativo para o padrão 802.16 de redes metropolitanas sem fio;
- A diferenciação de serviços através da incorporação de um escalonamento de pacotes baseado em mensagens e estações no protocolo proposto;
- A utilização de um esquema de alocação fixa no intervalo de reserva das estações clientes, para eliminar totalmente a possibilidade de colisões no acesso ao meio;
- Um modelo analítico para avaliação do protocolo proposto através de uma métrica de desempenho (mais especificamente, o tempo médio de espera na fila);
- A implementação do protocolo em uma ferramenta de simulação, visando a validação da modelagem analítica através dos resultados simulados.

1.5 Estrutura da Tese

O Capítulo 2 apresenta uma sucinta introdução aos conceitos inerentes a qualidade de serviço (QoS) em redes de computadores, mais precisamente, no ambiente de comunicação sem fio. Para isto, são descritas algumas técnicas utilizadas para escalonamento de pacotes, controle de admissão e policiamento de tráfego. Também são apresentados os diferentes tipos de tráfego suportados por aplicações em rede.

No Capítulo 3, uma breve descrição da topologia do IEEE 802.16 é apresentada. Além disso, é detalhado o funcionamento das camadas física e MAC, bem como, a arquitetura de QoS definida pelo padrão. Um comparativo entre os aspectos de QoS existentes nos padrões do IEEE para WLAN e WMAN é apresentado no final deste capítulo.

No Capítulo 4, alguns trabalhos relacionados a esta tese são apresentados. Primeiramente, é feita uma revisão bibliográfica dos principais protocolos de múltiplo acesso baseados em alocação dinâmica, encontrados na literatura. Posteriormente, algumas arquiteturas de QoS propostas na literatura para o padrão 802.16 são apresentadas.

O funcionamento das duas versões do protocolo proposto é descrito no Capítulo 5. Depois, são expostos alguns comentários a respeito das vantagens obtidas com a utilização no novo método de acesso.

O Capítulo 6 apresenta um modelo analítico para o tempo médio de espera na fila nas duas versões da solução proposta. Ainda neste capítulo, são descritas duas extensões para a Versão II do protocolo proposto, envolvendo prioridades variáveis entre as estações.

No Capítulo 7, os resultados numéricos obtidos através do modelo analítico descrito no capítulo anterior são apresentados. Além disso, este capítulo apresenta uma comparação entre os resultados obtidos analiticamente e por simulação.

O Capítulo 8 finaliza este trabalho consolidando os resultados apresentados no capítulo anterior através das conclusões e observações relevantes. As principais contribuições da tese são descritas também neste capítulo. Finalmente, algumas perspectivas para trabalhos futuros são sugeridas.

Por fim, no Apêndice A, os principais resultados referentes ao sistema de fila M/G/1 são apresentados.

Capítulo 2

Qualidade de Serviço

GARANTIR qualidade de serviço (QoS) em redes de computadores tornou-se uma necessidade básica devido aos novos tipos de aplicações utilizadas para transmissão de áudio e vídeo. Porém, esta tarefa é complexa, principalmente no contexto das comunicações sem fio. Nesse capítulo, os conceitos básicos inerentes a qualidade de serviço em redes de computadores são brevemente introduzidos. Em seguida, os diferentes tipos de tráfego envolvendo aplicações em rede são apresentados. Além disso, o capítulo descreve as principais técnicas utilizadas no escalonamento de pacotes, policiamento e controle de admissão para arquiteturas de QoS.

2.1 Visão Geral

Em alguns anos, houve um crescimento explosivo na utilização de aplicações que transmitem áudio e vídeo através de redes de computadores, as chamadas **aplicações multimídia** tais como, vídeos de entretenimentos, telefonia IP, rádio pela Internet, teleconferências, jogos virtuais, ensino à distância e outras. Porém, estes sistemas devem ser tratados de maneira diferente dos tradicionais programas que transmitem dados, como por exemplo, correspondência eletrônica (*e-mail*), páginas WWW, FTP, DNS, etc. Aplicações multimídia necessitam de um gerenciamento mais sofisticado do que sistemas de dados pois, a inerente natureza das redes baseadas no TCP/IP podem afetar a QoS oferecida ao usuário, visto que o TCP/IP não possui a capacidade de fornecer serviços diferenciados para tipos distintos de tráfego.

Em geral, aplicações multimídias são altamente sensíveis ao atraso fim-a-fim na transmissão de pacotes (*delay*) e variações neste atraso (*jitter*), mas podem tolerar algum nível de perda de dados. Por outro lado, aplicações tradicionais não são afetadas pelo atraso dos pacotes, mas a integridade dos dados transmitidos é de extrema importância. Por isso, a existência de algum mecanismo que trate estas diferentes necessidades nas transmissões em redes de computadores é imprescindível. Em [20], os autores identificam quatro princípios básicos para prover qualidade de serviço em aplicações multimídia:

1. Classificação de pacotes
2. Escalonamento e policiamento de tráfego
3. Alto índice de utilização de recursos
4. Controle de admissão

A classificação permite a distinção entre pacotes pertencentes a diferentes classes de tráfego e possibilita um tratamento diferenciado para cada pacote. Entretanto, simplesmente classificar os pacotes não garante que eles recebam um serviço com

a QoS desejada. A classificação é apenas um mecanismo para distingui-los. A diferenciação no tratamento destes pacotes é uma decisão do policiamento utilizado.

Idealmente, é desejável que haja um grau de isolamento entre fluxos distintos de tráfego, para que um mal comportamento de um determinado fluxo não afete os demais. Se um fluxo específico deve seguir certos critérios (como por exemplo, não exceder alguma taxa pré-estabelecida), um mecanismo de policiamento pode ser empregado para garantir que estes parâmetros serão observados. Uma outra alternativa para o isolamento do tráfego é a utilização de um escalonador de pacotes. Por exemplo, o escalonador pode alocar uma quantidade fixa da largura de banda do canal para cada fluxo.

Ao mesmo tempo que deseja-se prover o isolamento dos fluxos, é imprescindível utilizar eficientemente os recursos disponíveis, tais como, largura de banda e área de armazenamento (*buffer*). Para ilustrar este requisito, pode-se citar o exemplo onde é permitido a um determinado fluxo utilizar a largura de banda de outro fluxo que esteja inativo por um período de tempo.

Para aplicações que exigem um mínimo de QoS, deve ser permitida a utilização da rede (caso suporte a QoS necessária) ou bloqueada a “passagem” deste fluxo pela rede. Para isto, é necessário que o fluxo declare os seus requisitos de QoS. Este processo de declarar os requisitos de QoS e, aceitar ou rejeitar um fluxo na rede de acordo com estes requisitos, é chamado de controle de admissão.

Como foi dito no Capítulo 1, os sistemas de filas são utilizados para analisar o desempenho de redes que utilizam recursos compartilhados. Assim, gerenciar o comportamento destes sistemas de filas é o principal aspecto quando deseja-se garantir QoS para uma aplicação. As seções seguintes provêm uma visão geral de vários mecanismos de implementação dos quatro princípios listados acima, considerando as principais disciplinas de filas utilizadas.

2.2 Tipos de Tráfego

A atual infra-estrutura “de melhor esforço” das redes baseadas na arquitetura TCP/IP não atende as principais características, com relação ao atraso e perda, requeridas pelas aplicações multimídia. Contudo, um expressivo progresso tem sido obtido no sentido de especificar diferenciação de serviços para suportar estas aplicações. Dessa forma, a identificação do serviço, através da classificação de pacotes, é o primeiro passo para prover diferentes níveis de serviços.

Os pacotes que trafegam na rede podem ser divididos em três tipos básicos: voz, vídeo e dados. Voz e vídeo são exemplos de tráfego em tempo real, onde os bits são gerados periodicamente, formando um fluxo constante de dados. Se nenhum esquema de compressão é utilizado, este fluxo é chamado de tráfego com taxa constante de bits (*Constant Bit Rate* - CBR). Entretanto, esquemas de compressão convertem este tipo de tráfego para uma taxa variável de bits (*Variable Bit Rate* - VBR). Tráfego em tempo real não suporta grandes variações no atraso *jitter* durante as transmissões. Por outro lado, aplicações de dados, que não exigem tempo real, não possuem fortes restrições com relação ao atraso nas transmissões e, além disso, possuem taxa variável de bits (VBR) [21].

A utilização de modelos de tráfego para avaliação de desempenho em redes de computadores é de extrema importância. Em particular, a distribuição de Poisson tem sido bastante utilizada para esta finalidade [9]. Porém, no contexto das aplicações multimídia, onde existe a integração dos tráfegos de dados, voz e vídeo, modelos mais elaborados para caracterizar cada tipo de aplicação são necessários. Por exemplo, fontes de voz podem ser caracterizadas pelo modelo de fonte *on-off* e o modelo MMPP (*Markov Modulated Poisson Process*) pode ser utilizado para representar uma fonte de vídeo [21]. Já para o tráfego de dados, que possui uma natureza auto-similar, a distribuição de Pareto tem sido amplamente empregada.

As diversas aplicações existentes em redes de computadores necessitam serviços que podem ser genericamente avaliados em três dimensões [20]: perda de dados, largura de banda e atraso.

Perda de Dados

Algumas aplicações tais como, correio eletrônico, transferência de arquivos e transferência de documentos Web, etc, necessitam de uma transmissão completamente confiável, ou seja, sem nenhuma perda de dados. Por outro lado, aplicações multimídia como áudio ou vídeo em tempo real, podem suportar um certo nível de perda.

Largura de Banda

Da mesma forma, algumas aplicações necessitam de uma quantidade mínima de largura de banda para transmitir dados. Aplicações elásticas são aquelas que conseguem utilizar a quantidade de banda disponível, isto é, não são sensíveis a largura de banda. Certamente, quanto mais banda disponível, melhor.

Atraso

Como já foi dito anteriormente, aplicações de tempo real exigem rígidos requisitos de tempo na transmissão dos dados. Longos atrasos tornam estas aplicações menos “realísticas”. Contudo, mesmo para aplicações que não possuem estas exigências, pequenos atrasos são sempre melhores.

A Tabela 2.1 ilustra os requisitos de confiabilidade, largura de banda e atraso, de algumas das principais aplicações em redes de computadores [20].

Aplicação	Perda	Banda	Atraso
Transferência de arquivos	Sem perda	Elástica	Não
Correio eletrônico	Sem perda	Elástica	Não
Documentos Web	Sem perda	Elástica	Não
Áudio em tempo real	Tolerável	1Kbps - 1Mbps	Sim (mseg.)
Vídeo em tempo real	Tolerável	10Kbps - 5Mbps	Sim (mseg.)
Jogos interativos	Tolerável	1Kbps - 10Kbps	Sim (mseg.)

Tabela 2.1: Requisitos de algumas aplicações de redes (valores aproximados).

2.3 Escalonamento de Pacotes

O escalonamento de pacotes desempenha um papel crucial em sistemas que fornecem garantias de qualidade de serviço [20]. Dentro desse contexto, esta seção apresenta algumas das mais importantes disciplinas de escalonamento de pacotes estudadas na literatura: *first-in-first-out*, fila de prioridades e *weighted fair queuing*.

2.3.1 *First-In-First-Out* (FIFO)

A abordagem mais simples para gerenciar o escalonamento de pacotes é a disciplina *First-In, First-Out* (FIFO), onde todos os pacotes que chegam são colocados em uma fila comum e servidos pela ordem de chegada, como ilustra a Figura 2.1. Pacotes são descartados quando encontram a fila cheia (*packet loss*). Como a disciplina FIFO trata todos os pacotes de maneira igual, não é possível prover diferentes níveis de QoS para fluxos distintos. Além disso, sistemas FIFO estão sujeitos a *hogging* [22], que ocorre quando um usuário envia pacotes a uma alta taxa e ocupa todo o sistema, impedindo outros usuários de acessá-lo.

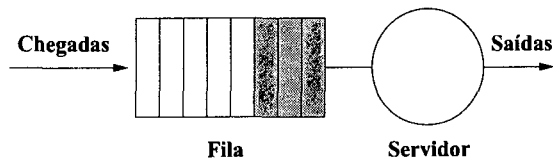


Figura 2.1: Abstração de um fila FIFO.

2.3.2 Fila de Prioridades

De acordo com esta disciplina, pacotes que chegam ao sistema são classificados em duas ou mais classes de prioridade, como mostra a Figura 2.2. Uma fila separada é mantida para cada classe. Esta disciplina transmite sempre o pacote que estiver na “cabeça” da fila de maior prioridade que não se encontra vazia. Por isso, esta disciplina é também conhecida como *Head-Of-the-Line* [10]. Para os pacotes que pertence a mesma classe de prioridade, uma disciplina FIFO pode ser utilizada.

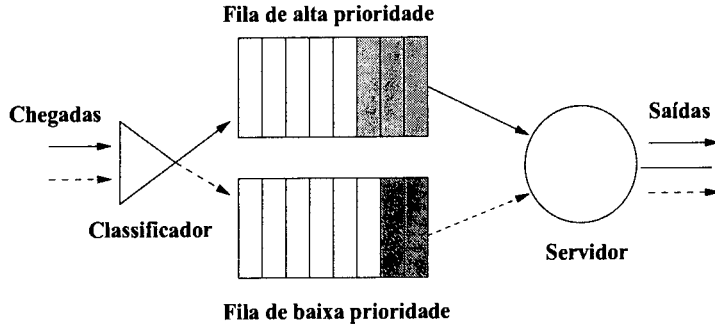


Figura 2.2: Modelo de fila com prioridades.

2.3.3 Weighted Fair Queuing (WFQ)

A disciplina WFQ [23, 24] é uma abstração generalizada da disciplina *round robin* [20], bastante utilizada em arquiteturas de QoS. Pacotes que chegam ao sistema são classificados e encaminhados para as filas correspondentes de cada classe, como ilustra a Figura 2.3. O escalonador serve as filas de maneira circular, ou seja, serve primeiro a classe de maior prioridade, depois a segunda maior prioridade e assim sucessivamente até a classe de menor prioridade, e então repete todo o processo. O que difere a disciplina WFQ da *round robin* é que a primeira tem a capacidade de aplicar uma quantidade diferenciada de serviço para uma determinada classe.

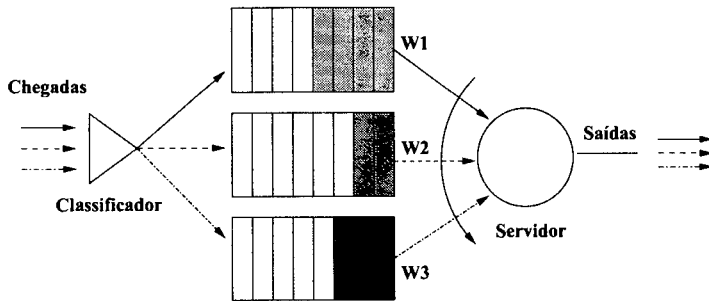


Figura 2.3: Weighted Fair Queuing (WFQ).

Na WFQ, para cada classe i é atribuída um “peso” w_i . Durante o intervalo de tempo em que existem pacotes da classe i para serem enviados, é garantida para esta classe uma fração do serviço igual a $w_i / (\sum w_j)$, onde a soma no denominador é dada por todas as classes que também possuem pacotes para enviar. Assim, se o canal de comunicação possui uma taxa de transmissão de C bps, a classe i sempre recebe no mínimo $C \cdot w_i / (\sum w_j)$ bps para transmissão de seus pacotes.

2.4 Controle de Admissão

A principal função do controle de admissão é decidir adequadamente se um canal de comunicação pode ou não aceitar uma nova conexão. Para isto, o controle de admissão deve aceitar ou rejeitar quando a fonte (um fluxo, uma estação, ou mesmo uma estação contendo vários fluxos) requisita uma nova conexão. Se a qualidade de serviço para todas as fontes que compartilham o mesmo canal (incluindo a nova) for satisfeita, a nova conexão será aceita. Caso contrário, essa conexão será rejeitada. Os índices de QoS podem ser expressos em termos de atraso máximo, probabilidade de perda, *jitter* e outras métricas de desempenho.

Para que este controle determine se a QoS no sistema será mantida, é necessário conhecer os diferentes tipos de fluxos de cada fonte. Para isto, cada fonte deve especificar o seu fluxo através de um conjunto de parâmetros conhecidos como **descritores de tráfego**. Estes descritores devem caracterizar os fluxos de tráfego de maneira compacta e eficiente. Baseado nos descritores de tráfego, o controle de admissão pode calcular a quantidade de banda que deve ser reservada para cada recurso.

Os mecanismos de controle de admissão podem ser classificados em não-estatísticos e estatísticos [25]. No primeiro caso, a alocação de recursos é realizada pela taxa de pico, ou seja, uma nova conexão será aceita se a soma das taxas de pico das conexões existentes e da nova conexão for menor que a capacidade do canal. Para o segundo caso, em geral aloca-se uma porção da banda entre a taxa de pico e a taxa média. O controle não-estatístico é de fácil implementação, porém, proporciona uma péssima utilização do canal. Em contra-partida, os mecanismos estatísticos alocam recursos de forma mais eficiente.

Existe uma grande variedade de mecanismos propostos na literatura para controle de admissão e para um estudo mais detalhado sobre o assunto recomenda-se [25]. Como exemplo de uma solução bastante utilizada, pode-se citar o conceito de **banda efetiva** [21] de cada fonte, que é um mecanismo de controle de admissão estatístico. O cálculo consiste em alocar uma quantidade de banda entre a taxa

média e a taxa de pico de uma determinada fonte.

2.5 Policiamento de Tráfego

Um vez que o controle de admissão aceite uma nova conexão, a qualidade de serviço será garantida se a fonte obedecer os descritores de tráfego que são especificados durante o estabelecimento desta conexão. Entretanto, se o fluxo de tráfego violar o “contrato” inicial, a rede poderá não suportar um desempenho aceitável. Assim, para impedir a violação dos contratos estabelecidos, deve existir algum mecanismo de policiamento do tráfego na rede.

O algoritmo de balde furado tem sido muito utilizado como mecanismo de policiamento de tráfego. Através dele, pode-se garantir: a taxa média de pacotes que um fluxo pode enviar na rede, a taxa de pico para este determinado fluxo e o tamanho máximo da rajada, ou seja, o número máximo de pacotes enviados em um curto período de tempo. Pode-se ainda, utilizar este mecanismo em conjunto com a disciplina de escalonamento WFQ vista anteriormente [20], para garantir um atraso máximo (d_{max}) a um determinado fluxo:

$$d_{max} = \frac{b_i}{C \cdot w_i / \sum w_j},$$

onde b_i é o tamanho máximo da rajada do fluxo i .

Capítulo 3

Padrão IEEE 802.16

ESTE capítulo apresenta uma visão geral do IEEE 802.16 e uma sucinta descrição da camada física e MAC, bem como, da arquitetura de QoS especificada pelo padrão. O escopo do padrão é especificar a interface aérea, incluindo a camada de acesso ao meio (MAC) e a camada física (PHY), para redes metropolitanas sem fio (WMAN), com diferenciação de serviços. Como será visto, o 802.16 fornece apenas o suporte para a implementação dessa arquitetura, possibilitando que cada fabricante introduza a sua própria solução. Por fim, é apresentado um comparativo entre os aspectos de QoS existentes nos padrões do IEEE para redes locais e metropolitanas sem fio.

3.1 Visão Geral

Em julho de 1999 o IEEE criou o grupo de trabalho 802.16, formado por integrantes das principais universidades e fabricantes, para desenvolver um padrão para sistemas BWA. A versão final foi aprovada em outubro de 2004 [5], com o seguinte escopo “especificar a interface aérea para sistema fixo de acesso sem fio à banda larga” [26]. Este sistema é conhecido também com interface aérea **IEEE WirelessMAN**. O 802.16 define como o tráfego sem fio é transmitido entre as estações clientes e uma estação base. Estes clientes podem ser usuários domésticos ou um centro comercial acessando a Internet, filiais de uma empresa conectadas a sua matriz, ou mesmo um Campus Universitário, como ilustra a Figura 3.1.

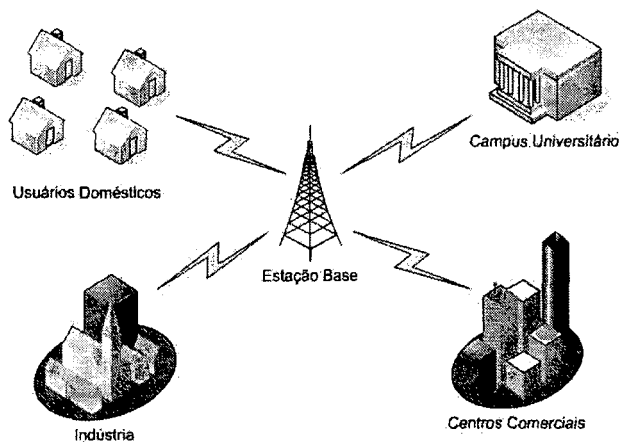


Figura 3.1: Variedade de estações clientes comunicando-se com uma estação base.

Esta tecnologia foi desenvolvida para alavancar o acesso sem fio à banda larga em redes metropolitanas (MANs), oferecendo desempenho comparável as tradicionais tecnologias de cabo e DSL. Entretanto, as principais vantagens do 802.16 são: a habilidade de prover serviços rapidamente, mesmo em áreas de difícil implantação de infra-estrutura; evitar gastos desnecessários com custos de instalações; a capacidade de ultrapassar limites físicos, como paredes ou prédios; alta escalabilidade; baixo custo de atualização e manutenção; dentre outros.

Sua arquitetura básica consiste de uma estação base (*Base Station* - BS) e uma ou mais estações clientes (*Subscriber Station* - SS), como mostra a Figura 3.2. A BS

é o nó central que coordena toda a comunicação e as SSs se localizam à diferentes distâncias da BS. Além disso, todo o tráfego de dados da rede passa pela BS, ou seja, não existe comunicação direta entre as SSs. A estação base pode estar conectada a uma outra infra-estrutura de rede (como por exemplo, a Internet), possibilitando uma extensão dos serviços oferecidos aos usuários. Da mesma forma, as estações clientes podem oferecer serviços diferenciados para usuários conectados através de uma rede local cabeada, ou sem fio.

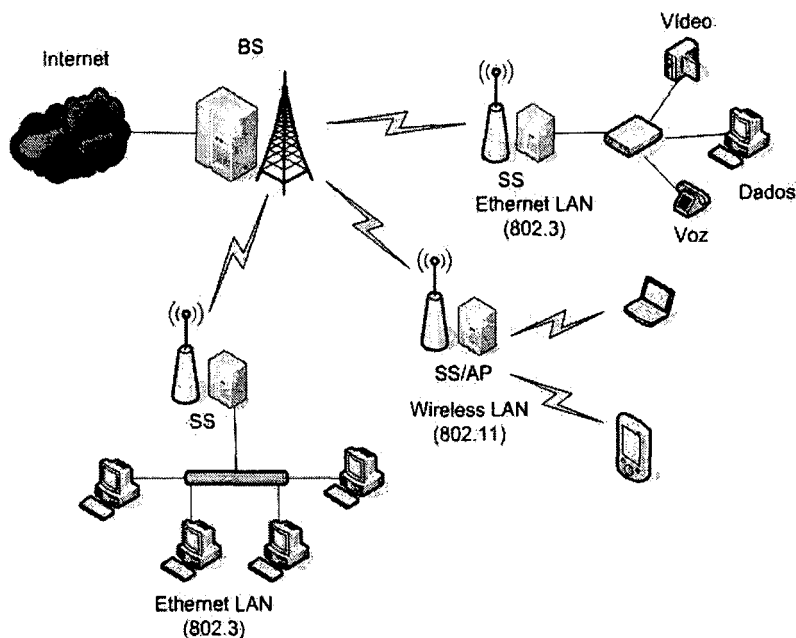


Figura 3.2: Arquitetura básica do sistema BWA.

3.2 Camada PHY e MAC

A camada física (PHY) do 802.16 opera em um intervalo de frequência que vai de 10 até 66 GHz. Além disso, dois novos protocolos para a camada física estão em fase de desenvolvimento. O padrão 802.16a, que opera em um intervalo de frequências entre 2 e 11 GHz, e o 802.16b, que utiliza a banda ISM (*Industrial, Scientific and Medical*) acima de 5 GHz. As taxas de transmissão de dados vão de 50 à 150 Mbps, dependendo da largura de frequência do canal e do tipo de modulação [26]. As transmissões ocorrem em dois canais diferentes: um canal de descida (*downlink* -

DL), com o fluxo de dados direcionado da BS para as SSs, e outro de subida (*uplink* - UL), com o fluxo de dados direcionado das SSs para a BS. No DL, os dados são transmitidos por difusão, enquanto no UL o meio é compartilhado através de múltiplo acesso.

O padrão fornece a flexibilidade de dois esquemas para alocação de banda: duplexação por divisão de freqüências (*Frequency-Division Duplexing* - FDD) e duplexação por divisão do tempo (*Time-Division Duplexing* - TDD). Basicamente, no FDD o DL e o UL utilizam freqüências diferentes, enquanto no TDD, os dois canais compartilham a mesma freqüência e os dados são transmitidos em tempos diferentes. O canal é segmentado no tempo e composto por quadros de tamanho fixo. Cada quadro é dividido em um sub-quadro para o DL e outro para o UL. A duração destes sub-quadros é dinamicamente controlada pela BS. A Figura 3.3 mostra a estrutura do quadro PHY com TDD.

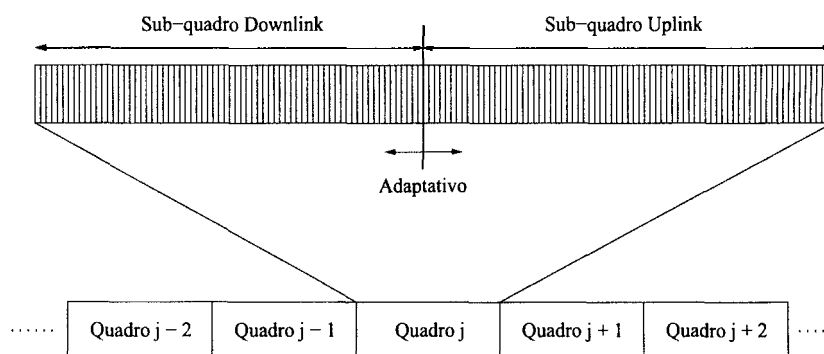


Figura 3.3: Estrutura do Quadro TDD

Durante o DL a transmissão é relativamente simples pois somente a BS transmite neste sub-quadro. Os pacotes de dados são transmitidos por difusão para todas as SSs, que por sua vez, capturam apenas os pacotes destinados a elas. Para o UL, a BS determina o número de segmentos que será atribuído para cada SS dentro do sub-quadro. Esta informação é transmitida por difusão pela BS através da mensagem UL-MAP no começo de cada quadro. A UL-MAP contém informações específicas (*Information Element* - IE) que incluem as oportunidades de transmissão, ou seja, os segmentos de tempo durante os quais a SS pode transmitir durante o sub-quadro UL. Após receber a UL-MAP, as estações transmitem os dados em segmentos de

tempo pré-definidos como indicados no IE. Na BS, é necessário um módulo de escalonamento do UL para determinar as oportunidades de transmissão (IEs) utilizando as requisições (BW-Request) enviadas pelas SSs. A Figura 3.4 ilustra a estrutura do quadro MAC no esquema de alocação TDD.

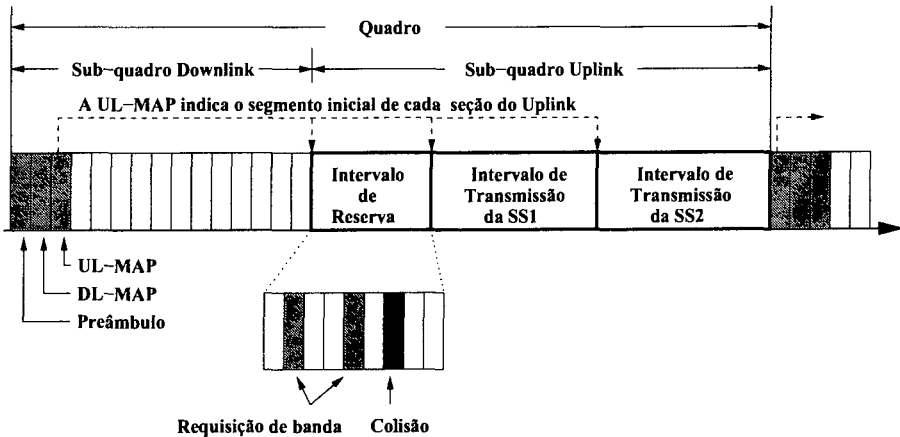


Figura 3.4: Estrutura do quadro MAC no esquema TDD.

As SSs [5] utilizam acesso aleatório e *piggybacking*¹ no sub-quadro UL para enviar requisições de oportunidades de transmissão para a BS. Esta é responsável por estabelecer um intervalo de reserva no início do UL para que as SSs possam requisitar as oportunidades de transmissões no próximo sub-quadro UL (ou em algum mais a frente, dependendo da ocorrência ou não de colisões). É importante notar que o 802.16 utiliza um protocolo de acesso ao meio baseado em alocação dinâmica, onde o período de reserva, que serve para identificar as demandas dos usuários, utiliza acesso aleatório. Depois de enviar a requisição de banda para a BS, a estação aguarda ser escalonada em algum sub-quadro UL mais a frente, como indica a Figura 3.5.

O padrão define o algoritmo *binary truncated exponential backoff* para resolução de colisões neste intervalo. Uma SS detecta a ocorrência de colisão caso a UL-MAP do próximo quadro não contenha nenhuma oportunidade de transmissão destinada a ela. Uma outra característica do padrão é o suporte a requisição de oportunidades de transmissão baseada em conexão (*Grants per Connection - GPC*) ou por estação (*Grants per Subscriber Station - GPSS*). Na GPSS, a estação requisita oportunidades de transmissão como um pacote para todos os serviços que ela mantém, e esta SS é

¹Requisições enviadas pelas SSs no final do quadro de dados, transmitidas durante o UL.

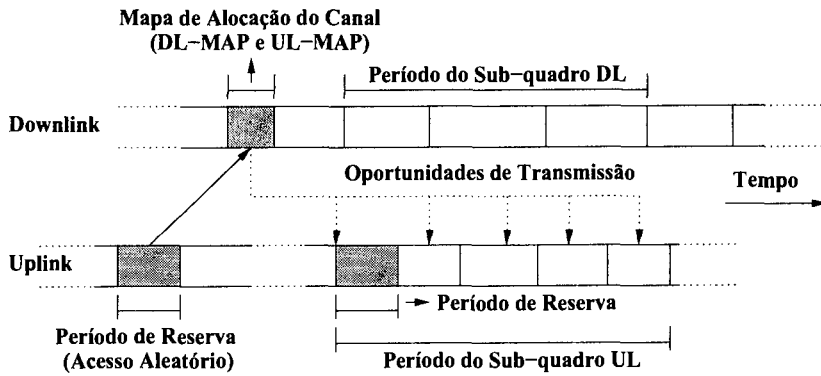


Figura 3.5: Estrutura de alocação do 802.16.

responsável por alocar as oportunidades recebidas entre os diferentes tipos de fluxos. Contudo, o 802.16 define apenas os mecanismos para sinalização de QoS, tais como BW-Request e UL-MAP, mas não define o escalonador do UL, ou seja, o mecanismo que determina as IEs na UL-MAP.

3.3 Arquitetura de QoS

A camada MAC do 802.16 define mecanismos de sinalização de QoS e funções para controlar a transmissão de dados entre a BS e as SSs. Dentro desse contexto, o padrão define quatro tipos de serviços associados a fluxos de tráfego [5], cada um com diferentes requisitos de QoS:

1. *Unsolicited Grant Service* (UGS): este serviço suporta tráfego com taxa constante (CBR) ou fluxos similares tais como, voz sobre IP (VoIP). Estas aplicações requerem uma constante alocação de banda.
2. *Real-Time Polling Service* (rtPS): este serviço é para aplicações de tempo real com taxa de transmissão variável (VBR) como por exemplo, MPEG vídeo ou teleconferência. Estas aplicações possuem requisitos específicos de banda, bem como, um atraso máximo tolerável.
3. *Non-Real-Time Polling Service* (nrtPS): este serviço é para fluxos sem requisitos de tempo real, mas que necessitam melhores condições do que os serviços

“de melhor esforço”, como por exemplo, transferência de arquivo. Estas aplicações são insensíveis ao atraso no tempo e requerem um mínimo de alocação de banda.

4. *Best Effort Service* (BE): este serviço é para tráfego “de melhor esforço”, onde não existe garantia de QoS, tais como HTTP. As aplicações recebem banda disponível após a alocação dos três fluxos anteriores.

No serviço UGS, BW-Request não é necessário. Para os demais tipos, o tamanho atual da fila é incluído no BW-Request para representar a demanda atual por banda de transmissão. Em resumo, o IEEE 802.16 especifica: o mecanismo de sinalização para troca de informações entre a BS e as SSs, como a configuração de conexões, BW-Request e UL-MAP; e o escalonamento do UL para serviço UGS. O padrão não define: o escalonamento do UL para serviços rtPS, nrtPS e BE; controle de admissão e o policiamento do tráfego.

A Figura 3.6 exhibe a arquitetura de QoS existente no 802.16. O módulo de escalonamento de pacotes do UL (*Uplink Packet Scheduling* - UPS) encontra-se na BS e controla todas as transmissões de pacotes no UL. Como o protocolo é orientado à conexão, a aplicação deve estabelecer uma conexão entre a BS e o fluxo de serviço associado (UGS, rtPS, nrtPS ou BE). A BS identifica as conexões com um CID (*Connection ID*) único para cada uma. O 802.16 define o processo de sinalização para o estabelecimento de uma conexão (*Connection Request, Response*) entre SS e BS, mas não especifica o processo de controle de admissão.

Todos os pacotes da camada de aplicação em uma SS são classificados de acordo com o CID e encaminhados para uma fila apropriada. A SS recupera o pacote na fila e transmite-o na rede no segmento de tempo determinado na UL-MAP enviada pela BS. A UL-MAP é definida pelo módulo UPS baseada nas mensagens BW-Request que reportam o tamanho atual da fila de cada conexão na SS.

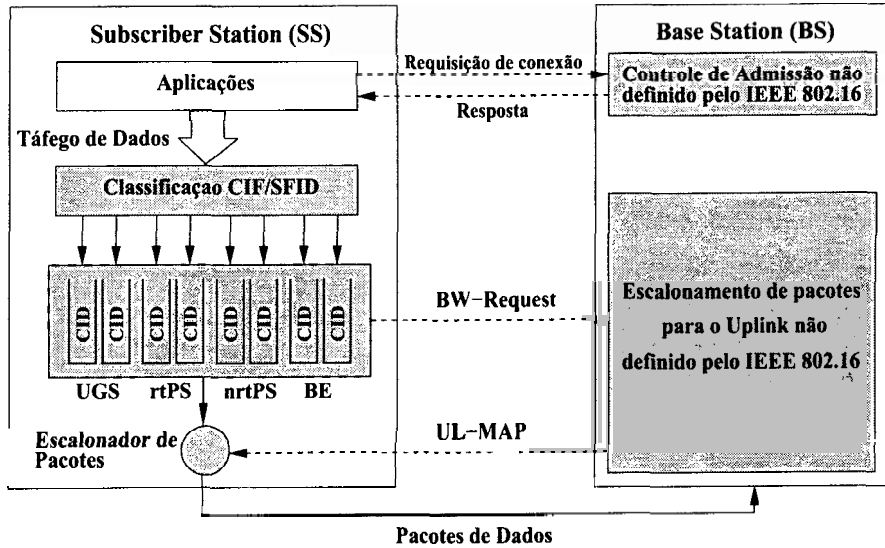


Figura 3.6: Arquitetura de QoS do IEEE 802.16.

3.4 Wi-Fi *versus* WiMAX

A principal característica de uma rede Wi-Fi é a sua simplicidade. Uma estação pode associar-se a um ponto de acesso ou a um *hot spot* de maneira simples e quase transparente para o usuário. Porém, esta simplicidade traz algumas limitações. Mesmo com as melhorias propostas pelo *draft* 802.11e, o Wi-Fi suporta apenas uma única conexão e parâmetros limitados de QoS [27]. O 802.11 é baseado em uma arquitetura distribuída, onde as operações na camada MAC são coordenadas entre os pontos de acesso e as estações.

Por outro lado, o WiMAX é baseado em uma arquitetura totalmente centralizada, onde a estação base tem o controle do acesso ao meio entre as estações sem fio. O 802.16 suporta múltiplas conexões que são completamente caracterizadas por vários parâmetros de QoS. Além disso, o 802.16 fornece uma classificação dos pacotes que permite mapear as conexões entre várias aplicações e interfaces distintas, como por exemplo, Ethernet, ATM, IP etc.

O IEEE 802.16 já “nasceu” com a habilidade de suportar diferentes níveis de serviços para tipos distintos de tráfego, incorporada naturalmente na camada MAC. Porém, o padrão define apenas uma arquitetura capaz de suportar QoS e não especifica uma solução completa para fornecer garantias ao serviço oferecido.

Capítulo 4

Trabalhos Relacionados

A TRAVÉS do capítulo anterior identifica-se que o 802.16 utiliza um protocolo de acesso ao meio baseado em alocação dinâmica com reserva. Neste capítulo, alguns trabalhos relacionados a protocolos de reserva propostos na literatura, são contextualizados em relação a redes metroplatinas sem fio. Em particular, os protocolos RPAC e MLMA são avaliados detalhadamente, por formarem a base da proposta descrita no próximo capítulo. Algumas arquiteturas de QoS propostas na literatura para sistemas BWA, também são apresentadas.

4.1 Protocolos de Reserva

Protocolos de acesso com alocação dinâmica utilizam informações de controle para coordenar as transmissões no meio de comunicação. É através deste controle que é estabelecida a ordem em que as estações acessam o meio, dependendo das suas demandas e do funcionamento do protocolo. A ordem de acesso pode ser determinada na forma de *polling* ou reservas explícitas por parte dos usuários. Além disso, este controle pode ser realizado de maneira centralizada ou distribuída.

Como foi visto no Capítulo 1, protocolos de acesso ao meio que utilizam um esquema de reserva proporcionam uma alocação dinâmica do canal de acordo com as demandas de cada usuário, evitando períodos de tempo ociosos sem transmissões. Além disso, pode-se evitar completamente a possibilidade de colisões entre mensagens que são enviadas ao mesmo tempo na rede.

A idéia básica dos protocolos de reserva é alocar parte da largura de banda do canal para as estações enviarem as suas requisições, informando a demanda necessária para transmitir suas mensagens de dados. Em geral, a atividade no canal pode ser vista como uma seqüência alternante de intervalos de reserva e de transmissão. Durante o período de reserva, também chamado de intervalo de escalonamento, vários algoritmos de acesso podem ser empregados para coordenar a transmissão dos pacotes de reserva. Seguindo este período, de acordo com as regras de escalonamento, é concedido acesso ao canal a cada estação que possui pacotes para transmitir.

Vários protocolos de acesso dinâmico foram propostos na literatura nas últimas décadas [13, 14, 12]. Em particular, os protocolos MSAP [28], BRAM [29], GSMA [30] e MLMA [31], possuem grande aplicabilidade para redes que cobrem grandes distâncias geográficas e possuem altas taxas de dados [12]. Dessa forma, estes protocolos podem ser facilmente empregados para controlar o acesso em redes metropolitanas, principalmente no contexto das comunicações sem fio. O protocolo RPAC [19] também aplica-se perfeitamente para este ambiente já que possui comportamento similar ao MLMA, como será visto ainda neste capítulo.

O protocolo MSAP (*Mini-Slotted Alternating Priorities*) funciona de maneira centralizada ou distribuída, porém, assume que todas as estações estejam na área de cobertura uma das outras. O mecanismo de acesso BRAM (*Broadcast Recognizing Access Method*) funciona de maneira totalmente descentralizada e também assume que as estações estejam na mesma área de cobertura. Já o protocolo GSMA (*Global Scheduling Multiple Access*), diferente dos dois anteriores, não implementa nenhuma política de prioridade para diferenciação de serviço. Dessa forma, estes protocolos não se adaptam adequadamente ao sistema BWA descrito no Capítulo 3, onde o controle é totalmente centralizado, as estações não estão necessariamente próximas umas das outras e a diferenciação de serviço é fundamental.

O protocolos MLMA e RPAC, por formarem a base da solução proposta no próximo capítulo, serão apresentados detalhadamente mais adiante. Antes disso, a próxima seção apresenta a notação matemática utilizada para avaliação de desempenho desenvolvida neste capítulo.

4.1.1 Notação Matemática

Os modelos matemáticos apresentados neste capítulo consideram filas para representar as estações da rede que são indexadas por i ($i = 1, 2, \dots, M$), onde M é o número total de terminais (filas). Além disso, os fluxos de tráfego pertencem a uma das P classes de prioridade, cada uma identificada pelo índice p ($p = 1, 2, \dots, P$), onde a classe 1 tem a maior prioridade e a classe P tem a menor prioridade.

O tráfego é representado por um processo de Poisson, onde λ_i^p representa a taxa de chegada de mensagens de classe p na estação i . Alguns modelos apresentados neste capítulo consideram tempo contínuo e outros consideram tempo discreto. Para tempo contínuo, a média e o segundo momento do tempo de serviço das mensagens de classe p na fila i são representados por b_i^p e $b_{2,i}^p$, respectivamente. No caso de modelos de tempo discreto, b_i^p e $b_{2,i}^p$ representam, respectivamente, a média e o segundo momento do número de pacotes contidos nas mensagens de classe p da estação i . Ainda em relação à tempo discreto, o canal é segmentado de forma que,

cada segmento é exatamente igual ao tempo de transmissão de um pacote.

O tráfego oferecido no canal é representado por:

$$\rho = \sum_{p=1}^P \sum_{i=1}^M \rho_i^p,$$

onde $\rho_i^p = \lambda_i^p b_i^p$. O tempo médio de espera na fila para mensagens de classe p na estação i é denominado \bar{W}_i^p .

4.1.2 MLMA

Em [32, 33], Werner Bux apresenta uma avaliação de desempenho para uma versão priorizada do protocolo de reserva MLMA (*Multi-Level, Multiple-Access*) [31], utilizando a teoria de renovação [9, 10]. A análise é realizada de forma genérica, podendo ser aplicada a qualquer protocolo onde a atividade no canal é vista como uma seqüência alternante de intervalos de transmissão e escalonamento, como mostra a Figura 4.1. Durante o intervalo de escalonamento, são definidas quais estações que possuem mensagens para enviar, acessam o canal no intervalo de transmissão do próximo quadro.

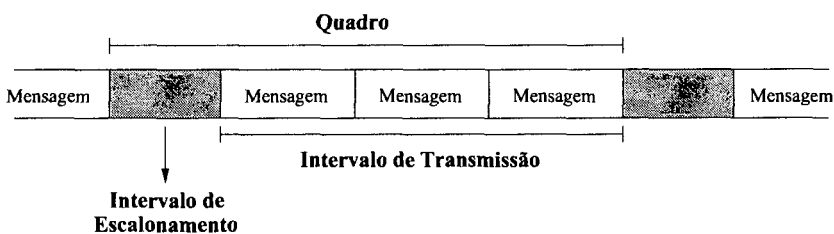


Figura 4.1: Esquema de acesso.

Com relação a ordem de transmissão, Bux considera três esquemas de escalonamento baseados em prioridade entre as estações: prioridade fixa, cíclica e complementar. A análise realizada desconsidera classes de tráfego distintas, ou seja, existe um único tipo de fluxo. Além disso, considera-se que a distribuição dos tempos de serviço são iguais em cada estação ($b_i = b$ e $b_{2,i} = b_2$; $i = 1, 2, \dots, M$).

Prioridade Fixa

Neste esquema, a ordem de serviço entre as M estações é fixa, ou seja, seguindo a fase de escalonamento, a fila 1 será servida, depois a fila 2 e assim sucessivamente, até a fila M . A ordem de transmissão entre mensagens da mesma fila segue a disciplina FIFO. Dessa forma, o tempo médio de espera na fila para as mensagens da estação i é representado pela equação 4.1:

$$\bar{W}_i = E[S] + \left(\frac{(1 + \rho_i)}{2} + \sum_{j=1}^{i-1} \rho_j \right) \frac{E[L^2]}{E[L]}, \quad (4.1)$$

onde $E[S]$ é o tempo médio do intervalo de escalonamento e, $E[L]$ e $E[L^2]$ representam, respectivamente, os dois primeiros momentos do tamanho do quadro:

$$E[L] = \frac{E[S]}{1 - \rho}; \quad \rho = \sum_{i=1}^M \rho_i < 1; \quad (4.2)$$

$$E[L^2] = \frac{E[S^2] - E[S]^2 + \lambda b_2 E[L]}{1 - \rho^2} + E[L]^2; \quad (4.3)$$

onde $\lambda = \sum_{i=1}^M \lambda_i$ é taxa agregada de chegada de mensagens de todas as estações.

Prioridade Cíclica

Nesta disciplina, a fila que tem prioridade i em um quadro assume a prioridade $(i + 1)$ no próximo. A estação de menor prioridade no quadro atual, obtém a maior prioridade no quadro seguinte. Para este tipo de escalonamento, o tempo médio de espera na fila é dado por:

$$\bar{W}_i = E[S] + \left(\frac{(1 + \rho_i)}{2} + \frac{1}{M} \sum_{j=2}^M (M - j + 1) \rho_{1+(i-j) \bmod(M)} \right) \frac{E[L^2]}{E[L]}, \quad (4.4)$$

com $E[L]$ e $E[L^2]$ dados pelas equações 4.2 e 4.3, respectivamente.

Prioridade Complementar

Esta disciplina define que uma estação i assume a prioridade i em um quadro, a prioridade $(M + 1 - i)$ no próximo e vice-versa. O tempo médio de espera para disciplina de escalonamento com prioridade complementar é dado por:

$$\bar{W}_i = E[S] + \left(\frac{1 + \rho}{2} \right) \frac{E[L^2]}{E[L]}; \quad i = 1, 2, \dots, M; \quad (4.5)$$

onde, novamente, os dois primeiros momentos do tamanho do quadro são dados pelas equações 4.2 e 4.3.

A Figura 4.2 ilustra o comportamento das três disciplinas, comparando o tempo médio de espera na fila para 11 estações. Assume-se que todos os terminais, exceto o terminal 6, geram a mesma quantidade de tráfego de 0,01 ($\rho_1 = \dots = \rho_5 = \rho_7 = \dots = \rho_{11} = 0,01$). Já o tráfego oferecido pela estação 6 é igual a 0,2, 0,5 e 0,8. Com isso, pode-se avaliar o comportamento das três disciplinas sob tráfegos de baixa, média e alta intensidade ($\rho = 0,3$, $\rho = 0,6$ e $\rho = 0,9$, respectivamente).

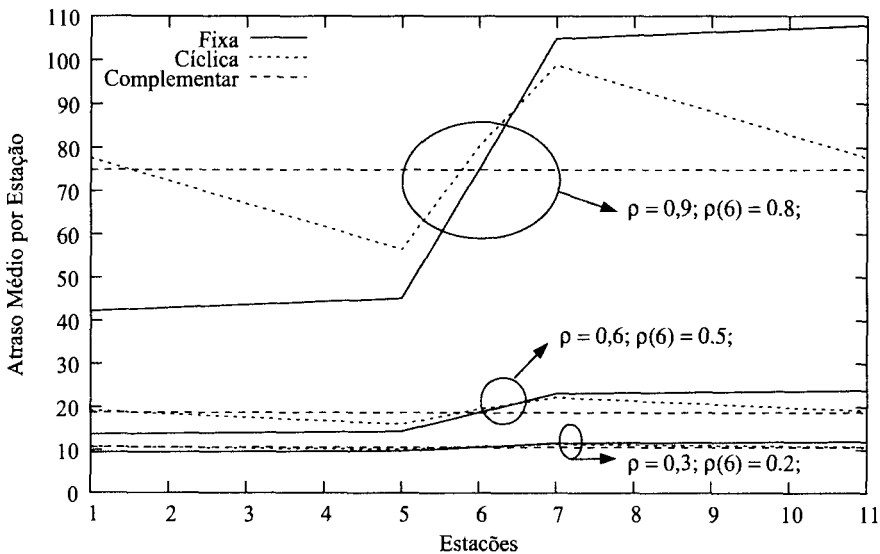


Figura 4.2: Comparação das disciplinas de prioridade: fixa, cíclica e complementar.

Pelo gráfico, pode-se notar que o atraso com a disciplina de prioridade fixa cresce a medida que diminui a prioridade no acesso ao meio, ou seja, com o aumento do índice das estações. Observa-se que esta “injustiça” no acesso ao meio pode ser

reduzida com a utilização do esquema cíclico, porém, não é totalmente evitada. As estações 5 e 7, por estarem mais próxima (em termos de prioridade) da estação 6, apresentam o menor e maior atraso médio, respectivamente. Para a disciplina de prioridade complementar, o atraso médio para todas as estações é o mesmo, independente da distribuição do tráfego entre elas. É válido ressaltar que, para um tráfego balanceado entre os terminais, a disciplina cíclica também proporciona o mesmo atraso médio entre as estações [32].

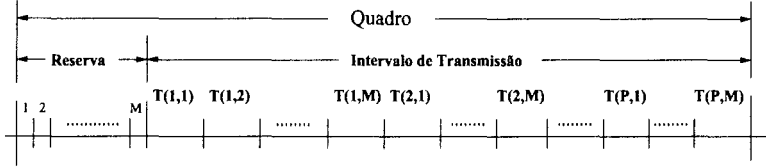
4.1.3 RPAC

Em [19, 34], é apresentado um protocolo de reserva denominado RPAC (*Reservation-Priority Access Control*), onde existem prioridades baseadas em mensagens (ou seja, tipos de fluxos) e estações. A estrutura geral é basicamente a mesma analisada em [32], porém, por introduzir regras de prioridades baseadas em mensagens, a modelagem analítica desenvolvida é mais geral. No RPAC é considerado um canal de comunicação por difusão e que todas as estações estão na mesma área de cobertura uma das outras. Além disso, considera-se tempo discreto (canal segmentado no tempo) e que o período de reserva é governado por uma disciplina TDMA (*Time-Division Multiple Acces*), com um único segmento reservado para cada estação requisitar o acesso ao meio, de acordo com as suas demandas. Com relação a regra de prioridade utilizada para determinar a ordem na qual as mensagens devem ser transmitidas, as duas versões do protocolo RPAC são consideradas:

RPAC I

Nesta versão, a priorização do acesso ao intervalo de transmissão é feita, primeiramente, entre as classes e depois entre as estações. Com isso, para quaisquer classes p e q ($\in \{1, \dots, P\}$), onde $p < q$, todas as mensagens de classe p são transmitidas antes de qualquer mensagem de classe q , independente da estação a qual pertença. Para mensagens pertencentes a mesma classe mas em estações distintas, a ordem de transmissão é de acordo com a ordem na qual as estações acessam o canal (primeiro

a estação 1 e por último a estação M). Para mensagens na mesma estação com a mesma classe de prioridade, as transmissões ocorrem por ordem de chegada (FIFO). O comportamento do RPAC I está ilustrado na Figura 4.3.



$T(p,i)$ = intervalo de tempo reservado para transmissão das mensagens de classe p da estação i ;

Figura 4.3: Comportamento do RPAC I.

A análise desenvolvida em [19] expressa o tempo médio de espera na fila para as mensagens de classe p na estação i para a versão I do RPAC através da equação 4.6, assumindo um intervalo de reserva com M segmentos (TDMA com um segmento para cada estação).

$$\bar{W}_i^p = M + \left(\frac{(1 + \rho_i^p)}{2} + \sum_{j=1}^{p-1} \sum_{g=1}^M \rho_g^j + \sum_{j=1}^{i-1} \rho_j^p \right) \frac{E[L^2]}{E[L]} - \frac{1}{2} \quad (4.6)$$

$E[L]$ e $E[L^2]$ representam, respectivamente, o primeiro e segundo momento do tamanho do quadro, e são obtidos através das equações 4.7 e 4.8.

$$E[L] = \frac{M}{1 - \rho}; \quad \rho = \sum_{p=1}^P \sum_{i=1}^M \rho_i^p < 1. \quad (4.7)$$

$$E[L^2] = \frac{1}{1 - \sum_{p=1}^P \sum_{i=1}^M (\rho_i^p)^2} \left\{ M^2 + \left[2M \sum_{p=1}^P \sum_{i=1}^M \rho_i^p + \sum_{p=1}^P \sum_{i=1}^M \lambda_i^p b_{2,i}^p \right] E[L] + \right. \\ \left. + \left[\sum_{p=1}^P \sum_{i=1}^M \sum_{\substack{j=1 \\ j \neq i}}^M \rho_i^p \rho_j^p + \sum_{p=1}^P \sum_{\substack{q=1 \\ q \neq p}}^P \sum_{i=1}^M \sum_{k=1}^M \rho_i^p \rho_k^q \right] E[L]^2 \right\} \quad (4.8)$$

RPAC II

A segunda versão do protocolo RPAC prioriza, primeiro, o acesso das estações e depois as classes de tráfego dentro de cada estação. Assim, para quaisquer terminais i e j ($\in \{1, \dots, M\}$), onde $i < j$, todas as mensagens no terminal i são transmitidas antes de qualquer mensagem no terminal j , independente da sua classe de prioridade. Em qualquer terminal, as mensagens são transmitidas de acordo com suas prioridades e em ordem de chegada, caso pertençam a mesma classe, ou seja, em cada terminal, a disciplina de prioridade HOL é aplicada com a maior prioridade atribuída para a classe 1 e a menor para a classe P . A Figura 4.4 mostra o comportamento da segunda versão do RPAC.

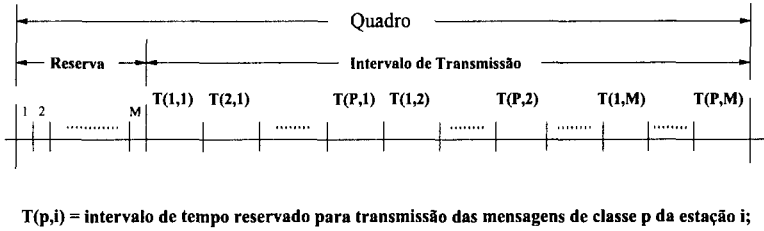


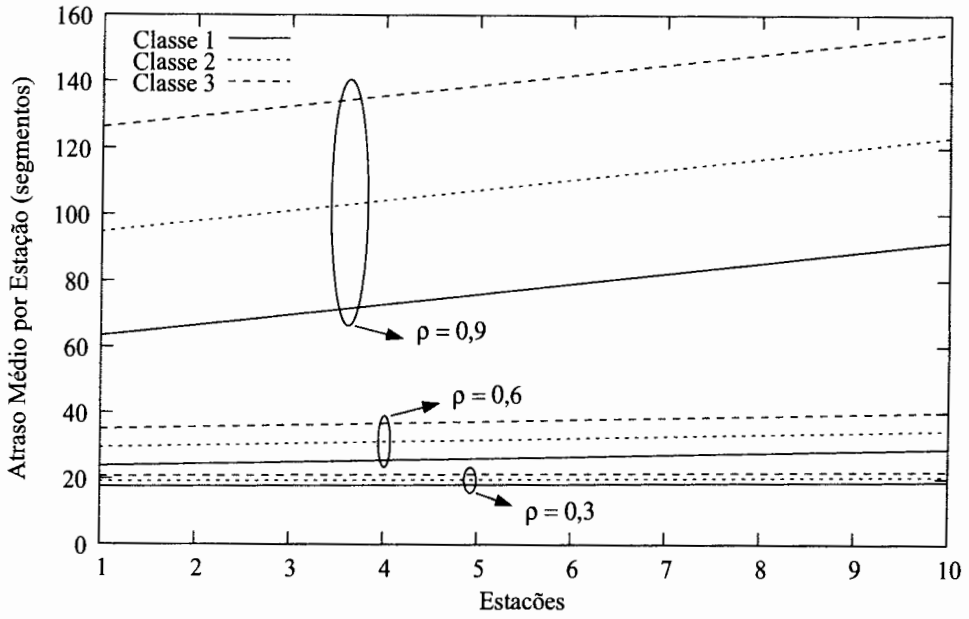
Figura 4.4: Comportamento do RPAC II.

Da mesma forma, [19] apresenta uma expressão fechada para o tempo médio de espera na fila para as mensagens de classe p na estação i , através da equação 4.9:

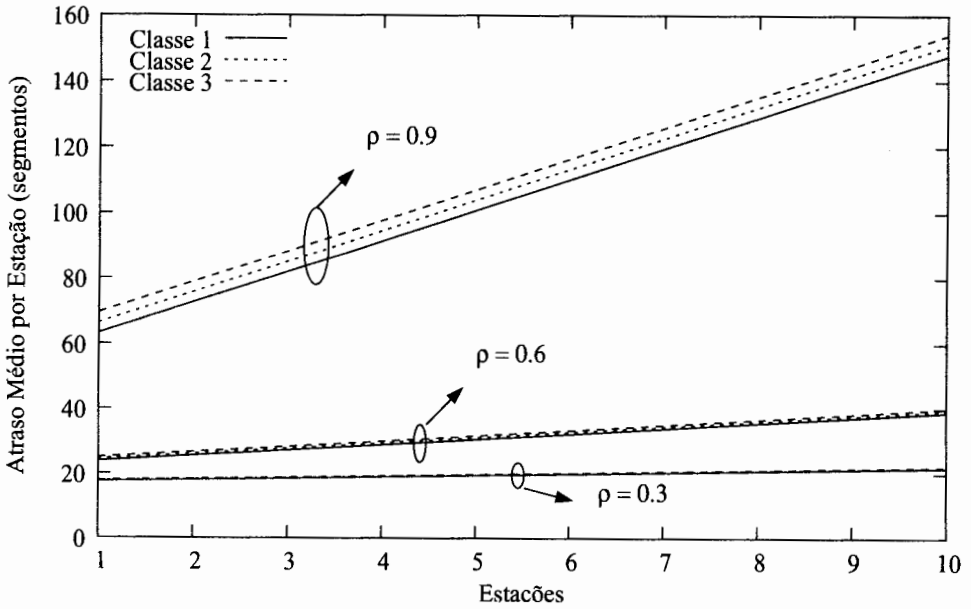
$$\bar{W}_i^p = M + \left(\frac{(1 + \rho_i^p)}{2} + \sum_{j=1}^{i-1} \sum_{k=1}^P \rho_j^k + \sum_{j=1}^{p-1} \rho_i^j \right) \frac{E[L^2]}{E[L]} - \frac{1}{2}; \quad (4.9)$$

onde $E[L]$ e $E[L^2]$ também são dados pelas equações 4.7 e 4.8, respectivamente.

A Figura 4.5 retrata o comportamento das duas versões do RPAC, comparando o tempo médio de espera na fila para 10 estações e 3 classes de tráfego. O tráfego é balanceado entre as estações e os valores são obtidos para três níveis de intensidade de tráfego: baixo ($\rho = 0,3$), médio ($\rho = 0,6$) e alto ($\rho = 0,9$). Pela Figura 4.5(a), pode-se notar a diferenciação obtida entre as 3 classes de tráfego, que é ainda mais intensa para uma alta taxa de tráfego. Esta diferenciação é menor para versão II do RPAC, como mostra a Figura 4.5(b), porque a priorização é feita primeiro por



(a) RPAC I



(b) RPAC II

Figura 4.5: Tempo médio de espera por estação: RPAC I (a) e RPAC II (b).

estações e não por classe. Além disso, o RPAC II promove uma “injustiça” no acesso ao meio entre os terminais de forma que, aumenta o tempo médio de espera na fila para as mensagens das estações de baixa prioridade.

4.2 Qualidade de Serviço em 802.16

Vários algoritmos de escalonamento e arquiteturas de QoS para BWA, têm sido propostos na literatura [35, 36, 37, 38, 39, 40, 41], já que o padrão fornece apenas mecanismos de sinalização e não especifica nenhum algoritmo para escalonamento e controle de admissão. Contudo, por ser um padrão relativamente novo¹, muitas destas soluções abordam apenas a implementação/adição de uma nova arquitetura de QoS incorporada ao 802.16. Poucos trabalhos têm sido apresentados envolvendo alterações no protocolo MAC do padrão, para que o tráfego escalonado possa ser eficientemente escoado [42]. Além disso, nenhum modelo analítico foi apresentado, buscando representar de forma exata alguma métrica de desempenho do protocolo de acesso (como atraso ou vazão).

Em [35], GuoSong Chu *et al.* propõem uma arquitetura de QoS para o 802.16, onde a BS utiliza apenas o método de requisições por estações GPSS (vide Seção 3.2). Com isso, a arquitetura proposta necessita a inclusão de um escalonador em cada SS, para tratar os diferentes fluxos de tráfego. Nas SSs, o escalonador aplica a disciplina WFQ para fluxos UGS e rtPS; *round robin* para nrtPS e FIFO para BE (vide Capítulo 2, para detalhes sobre estas disciplinas de escalonamento). Além da necessidade de um escalonador em cada SS, a arquitetura proposta não foi avaliada através de resultados analíticos ou simulados.

Em [36, 37, 38], Mohammed Hawa *et al.* especificam uma arquitetura de escalonamento para os padrões DOCSIS e 802.16. Para isso, o escalonador mantém três tipos de fila com diferentes disciplinas de serviço. A fila do tipo 1, que serve para os fluxos de serviços UGS, executa uma disciplina de serviço FIFO. As filas do

¹A última versão do padrão é de outubro de 2004 [5].

tipo 2 e 3, que são utilizadas para os demais tipos de fluxos (rtPS, nrtPS e BE), executam, respectivamente, as disciplinas FIFO e fila de prioridade. O escalonador utiliza a disciplina WFQ entre as filas do tipo 2 e 3. Além disso, é proposto um novo algoritmo para minimizar os efeitos do acesso aleatório no período de reserva, como alternativa ao *binary exponential backoff* empregado pelos padrões. Este algoritmo busca encontrar o melhor tamanho para a janela de contenção, visando diminuir a probabilidade de colisão entre as requisições dos usuários.

Em [39], Aura Ganz *et al.* propõem um escalonador de pacotes para o subcanal UL do 802.16, baseado em uma estrutura hierárquica de filas. Seguindo o escalonador proposto, a alocação de banda entre fluxos distintos segue uma disciplina de prioridade fixa, da maior (fluxo UGS) para a menor (fluxo BE). A alocação de banda entre fluxos iguais segue diferentes disciplinas de serviço. Para conexões UGS, o canal é alocado em quantidade fixa, dependendo da demanda total das conexões. A alocação de banda para conexões rtPS segue uma disciplina de serviço EDF (*Earliest Deadline First*), onde o pacote com menor *deadline* (“tempo de vida”) é transmitido primeiro. Conexões nrtPS são servidas de acordo com a disciplina WFQ. E, por fim, o restante da banda é alocado igualmente entre as conexões BE. Os autores desenvolveram um modelo de simulação para avaliar o comportamento do escalonador proposto. Contudo, além de apresentar apenas resultados simulados, os autores desconsideram a complexidade de implementação desta solução hierárquica e não definem claramente como é feita a requisição de banda.

Em [40], Dong-Hoon Cho *et al.* propõem uma nova arquitetura de QoS para 802.16 onde o escalonamento é baseado no tempo de vida do pacote de cada tipo de fluxo. Para isto, os autores aplicam o conceito de *arrival-service curve* para determinar o tempo de chegada e o tempo de vida de cada pacote. Além disso, os autores declaram através de uma análise matemática que, o melhor tamanho para a janela de *backoff* afim de evitar colisões durante o período de reserva é igual ao número de estações ativas na rede. Contudo, não foi proposto nenhum método novo para o intervalo de reserva e o trabalho não especifica claramente como é calculado o tempo de vida de cada pacote.

Em [41], Sung-Min Oh *et al.* declaram que o desempenho do IEEE 802.16 é bastante afetado pelo tamanho do intervalo de reserva (que é baseado em contenção), devido à probabilidade de colisões durante este período. O trabalho apresenta uma análise estocástica para encontrar o melhor tamanho para o período de reserva, com o intuito de otimizar a utilização do meio. Através desta análise, foi constatado que o melhor tamanho para o período de reserva é duas vezes o número de usuários ativos na rede. É válido observar que este resultado difere daquele exposto em [40], onde o melhor tamanho para a janela de contenção é igual ao número de estações ativas.

Novamente, é importante notar que nenhuma das propostas apresentadas nesta seção envolvem alterações no próprio protocolo de acesso ao meio. Além disso, algumas necessitam da adição de algum componente de hardware e/ou camada de software para especificar uma solução para o escalonamento de pacotes. Como será visto no próximo capítulo, a solução proposta neste trabalho busca introduzir um mecanismo de escalonamento incorporado ao protocolo MAC do 802.16. Dessa forma, pode-se diferenciar tipos distintos de tráfego de maneira rápida e eficiente, proporcionando um primeiro passo para especificação de uma arquitetura completa de QoS para redes metropolitanas sem fio.

Capítulo 5

Protocolo Proposto

ESTE capítulo descreve o funcionamento do protocolo de reserva proposto para controlar o acesso no sub-canal de *uplink* em sistemas BWA. Como será visto, o protocolo proposto incorpora um escalonador de pacotes baseado em mensagens e/ou estações. Duas versões distintas são apresentadas. Uma versão baseada em mensagens e outra baseada em estações.

5.1 Descrição

Como foi visto no Capítulo 3, a camada MAC do IEEE 802.16 utiliza um protocolo de acesso dinâmico, onde o esquema de reserva para o sub-canal UL é baseado em acesso aleatório. Além disso, o padrão não define como os diferentes tipos de tráfego serão escalonados no sub-canal UL. Dentro desse contexto, este capítulo descreve um protocolo alternativo para WMAN, onde o esquema de reserva é baseado em um controle fixo e o escalonamento do tráfego é realizado de maneira simples, baseado em mensagens e/ou estações. Visto que no sub-canal DL não existe múltiplo acesso, ou seja, somente a BS transmite dados, este trabalho preocupa-se apenas em propor uma forma de escalonamento para os diferentes tipos de tráfego no sub-canal UL.

O protocolo proposto é o próprio RPAC (*Reservation-Priority Access Control*) descrito em [19, 34], adaptado para o 802.16, onde o período de reserva é governado pelo esquema TDMA com um segmento reservado para cada estação na rede de forma fixa. Após este período, as estações transmitem suas mensagens de acordo com as regras de prioridade estabelecidas. No RPAC é considerado um canal de comunicação por difusão onde todas as estações estão na mesma área de cobertura uma das outras, não existindo portanto um controle centralizado do acesso ao meio através da figura de uma estação base.

A principal vantagem da utilização do TDMA (alocação fixa) para requisição de banda é garantir a reserva do canal de forma simples e eficiente. Por outro lado, a utilização de uma disciplina de *polling* ou *probing* para o período de reserva, exige múltiplos chaveamentos dos dispositivos sem fio do modo de transmissão para o modo de recepção e vice-versa, no sub-quadro UL. Ou ainda, que as estações estejam na área de cobertura umas das outras [43]. E como desvantagem tem-se o desperdício do canal com segmentos alocados para estações que não possuem mensagens para transmitir. Porém, este desperdício é relativamente pequeno (principalmente para tráfego de média à alta intensidade de carga [44]) se comparado ao tempo que uma mensagem pode ficar armazenada no *buffer* devido à ineficiência do algoritmo para resolução de colisões [45].

O protocolo proposto também utiliza o TDMA para requisição de banda pelas SSs, onde é determinado de maneira fixa um segmento para cada estação na rede. Porém, diferente do RPAC, estas requisições são processadas de maneira centralizada pela BS, que coordena o acesso ao canal de transmissão. Com isso, não existe a limitação de todas as estações estarem na mesma área de cobertura uma das outras, visto que as SSs devem estar apenas dentro do alcance da BS. Outra característica relevante desse protocolo é a incorporação de um escalonador de tráfego que utiliza regras de prioridade, possibilitando assim, o suporte à QoS através da diferenciação de serviços baseados em mensagens e estações, como será visto adiante.

Após a apresentação das características do protocolo, será descrito em detalhes o seu funcionamento. Considera-se um sistema em que o canal de comunicação é totalmente sincronizado pela BS através da segmentação do tempo em intervalos de duração fixa de τ segundos, onde as transmissões só ocorrem no começo de cada segmento. No intervalo de reserva é atribuído um segmento de tempo a cada estação para requisição de banda. A estação envia para a BS no seu segmento pré-alocado todas as informações necessárias (descritores de tráfego) para transmitir as mensagens que estão armazenadas no seu *buffer*. Neste trabalho só será abordado o esquema de alocação TDD, porém os resultados apresentados podem ser facilmente estendidos para o esquema FDD.

A estrutura do quadro MAC para o protocolo proposto está ilustrada na Figura 5.1. Diferente do padrão 802.16, propõem-se um canal onde os quadros não possuem tamanho fixo. Basicamente, o tamanho do quadro depende da quantidade de pacotes que chegaram no quadro anterior. No período de reserva, que localiza-se no final do sub-quadro UL e não no início como no 802.16, a estação informa para a BS os tipos dos serviços para os quais está requisitando banda e a quantidade de pacotes que chegaram para cada serviço. Após processar todas as requisições, a BS envia, no sub-quadro DL do próximo quadro, a UL-MAP com as oportunidades de transmissões para todas as estações.

Para efeito de análise do protocolo, a atividade no canal pode ser vista como uma seqüência de intervalos de reserva, *downlink* e de transmissão onde cada seqüência

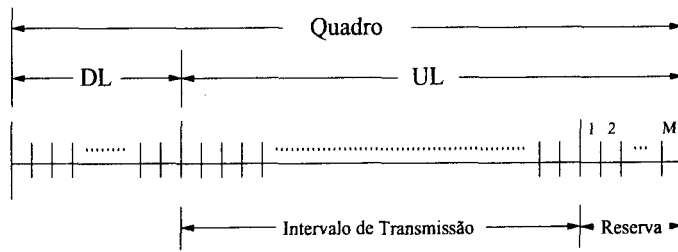


Figura 5.1: Estrutura do quadro MAC para o protocolo proposto.

constitui um **ciclo de transmissão**, como ilustrado na Figura 5.2. É importante notar que existe uma diferença entre o quadro MAC e o ciclo de transmissão, apesar de ambos terem um tamanho igual pois o intervalo de reserva é fixo. Na realidade, o n -ésimo ciclo é formado pelo intervalo de reserva do $(j - 1)$ -ésimo quadro, mais o intervalo de *downlink* e o intervalo de transmissão (parte do intervalo de *uplink*) do j -ésimo quadro. Com esta definição dos ciclos de transmissões pode-se utilizar a abordagem analítica descrita em [19] para análise do tempo de espera das mensagens, como será visto no Capítulo 6. Pela figura, pode-se definir L_n^R como o tamanho do intervalo de reserva do n -ésimo ciclo, L_n^{DL} como o tamanho do *downlink* do n -ésimo ciclo e L_n^T como o tamanho do n -ésimo intervalo de transmissão, todos medidos em segmentos. Assim, $L_n = L_n^R + L_n^{DL} + L_n^T$ é o tamanho total do n -ésimo ciclo. Seguindo o protocolo TDMA, cada intervalo de reserva é composto de M segmentos ($L_n^R = M; n = 1, 2, 3, \dots$), onde M é o número de estações na rede. Neste período, que tem a duração de $M\tau$ segundos, cada estação é associada a um segmento de maneira fixa.

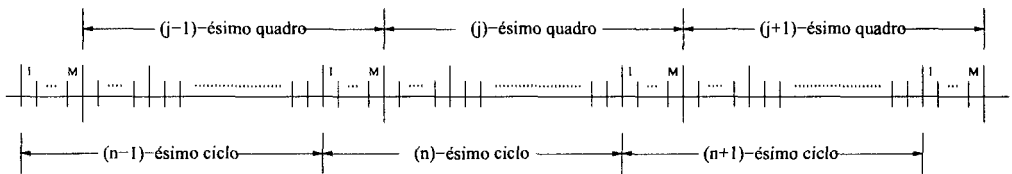


Figura 5.2: Ciclos consecutivos de transmissão.

O tamanho do ciclo atual depende da quantidade de mensagens que chegaram no ciclo anterior. Por exemplo, o tamanho do n -ésimo ciclo depende da quantidade de pacotes que chegaram nas mensagens do $(n - 1)$ -ésimo ciclo. Isto acontece porque a requisição para as mensagens que chegaram durante o $(n - 1)$ -ésimo ciclo será

transmitida no período de reserva do n -ésimo ciclo. Após este período, a BS realiza o processamento centralizado das oportunidades de transmissões e envia a UL-MAP no sub-canal DL, ainda no n -ésimo ciclo. Depois, as estações transmitem suas mensagens no intervalo de transmissão do mesmo ciclo, seguindo as prioridades estabelecidas na UL-MAP. Então, mensagens chegando durante o ciclo corrente são transmitidas somente no ciclo subsequente.

Propõem-se um protocolo de acesso ao meio com prioridades baseadas em mensagens ou em estações, em conformidade com o protocolo 802.16 que utiliza admissões GPC ou GPSS. Adotou-se que, seguindo a fase de reserva e sub-quadro DL, o canal é alocado para as estações seguindo a seqüência $1, 2, 3, \dots, M$. Assim, de acordo com as regras de prioridades utilizadas para determinar a ordem na qual as mensagens devem ser transmitidas durante o intervalo de transmissão, as seguintes versões do protocolo são definidas:

- **Versão I**, sob a qual, para quaisquer $p, q \in \{1, \dots, P\}$ tal que $p < q$, todas as mensagens de classe p são transmitidas antes de qualquer mensagem de classe q , independente da estação a qual pertença. Para mensagens pertencentes a mesma classe mas em estações distintas, a ordem de transmissão é de acordo com a ordem na qual as estações acessam o canal (primeiro a estação 1 e por último a estação M). Para mensagens na mesma estação com a mesma classe de prioridade, as transmissões ocorrem por ordem de chegada.
- **Versão II**, sob a qual, para quaisquer $i, j \in \{1, \dots, M\}$ tal que $i < j$, todas as mensagens no terminal i são transmitidas antes de qualquer mensagem no terminal j , independente da sua classe de prioridade. Em qualquer terminal, as mensagens são transmitidas de acordo com suas prioridades e em ordem de chegada caso elas pertençam a mesma classe, ou seja, em cada terminal, a disciplina de prioridade HOL é aplicada com a maior prioridade atribuída para a classe 1 e a menor para a classe P .

O comportamento do canal de acordo com as versões I e II do protocolo proposto

está ilustrado nas Figuras 5.3 e 5.4 respectivamente. Nota-se que, na Versão II um terminal de alta prioridade transmite todas as suas mensagens antes de um terminal de baixa prioridade. Então, diferente da Versão I, é possível que mensagens de baixa prioridade sejam transmitidas antes de mensagens com maiores prioridades.

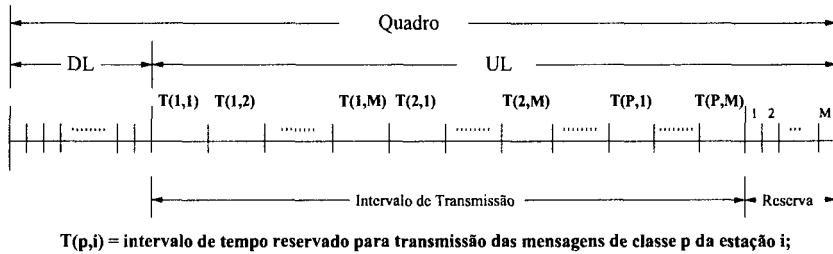


Figura 5.3: Comportamento da Versão I do protocolo proposto.

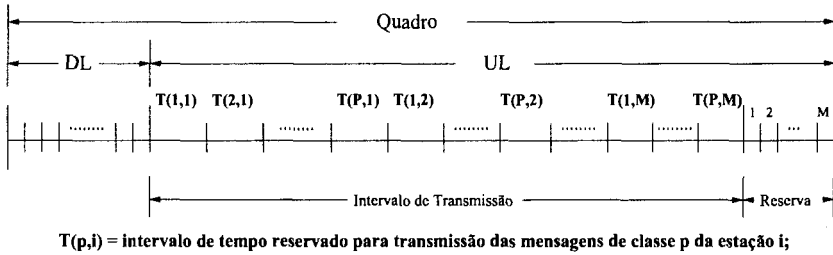


Figura 5.4: Comportamento da Versão II do protocolo proposto.

5.2 Comentários

Como visto na seção anterior, o protocolo proposto incorpora funções de escalonamento para diferentes tipos de fluxos. Esta habilidade permite a diferenciação de serviços baseada na classificação do tráfego escoado. Dessa forma, aplicações multimídia podem ser tratadas de maneira diferente das aplicações de dados. Visto que o padrão IEEE 802.16 já especifica a classificação dos pacotes, com a utilização do método de acesso proposto, dois dos quatro princípios básicos para garantir QoS são abordados.

Porém, observa-se que é necessário a existência de um rígido controle de admissão realizado pela BS e também pelas SSs, para que um tráfego elevado de uma determinada classe (ou estação) não sobrecarregue o canal e afete o tempo de resposta

das demais. Com isso, uma arquitetura completa para fornecer QoS em sistemas BWA pode ser aplicada, utilizando-se a técnica de banda-efetiva para controlar o processo de admissão de conexões e o algoritmo de balde furado para o policiamento do tráfego.

Uma outra vantagem do protocolo proposto é a utilização do TDMA para o período de reserva das estações clientes. Dessa forma, consegue-se eliminar totalmente a possibilidade de colisões de pacotes, aumentando o fator de utilização da rede (principalmente para altas intensidades de tráfego).

Capítulo 6

Modelagem Analítica

PARA avaliar o desempenho do protocolo proposto, este capítulo apresenta uma modelagem analítica para o tempo médio de espera na fila das mensagens transmitidas no sub-canal UL, nas duas versões descritas no capítulo anterior. Esta modelagem baseia-se em tráfego poissoniano e *buffer* infinito tanto na BS quanto nas SSs. Primeiramente, será detalhada a análise para a Versão I do protocolo proposto e em seguida, uma adaptação do resultado obtido para a Versão II é apresentada. Por fim, são incorporadas a Versão II do protocolo duas extensões envolvendo prioridades variáveis entre as estações.

6.1 Modelagem da Versão I

A técnica utilizada para obter o tempo médio de espera para as mensagens é similar ao método empregado em [19, 46, 47]. Para maiores detalhes, recomenda-se a leitura destas referências. Além disso, as duas extensões aplicadas para a Versão II do protocolo proposto baseiam-se no esquema de prioridades variáveis proposto por [32].

Considerando um sistema com uma BS e M ($M \geq 1$) estações clientes (SSs), todas já associadas com a estação base, onde cada estação possui *buffer* infinito. O canal de transmissão tem uma taxa de C bits/s e é considerado sem erro. As mensagens geradas em cada estação são compostas de unidades fixas de dados (pacotes) e o tempo de transmissão de cada pacote é igual a um segmento de tempo (τ). As mensagens são compostas por um número aleatório de pacotes, onde cada pacote contém μ^{-1} bits. Como o tempo de transmissão de um pacote é considerado igual ao tempo de um segmento, temos que $\tau = (\mu C)^{-1}$ (exatamente um segmento para transmitir um pacote).

As mensagens que chegam em cada estação pertencem a uma das P diferentes classes. Assume-se que a mensagem da classe 1 possui uma maior prioridade e a mensagem da classe P possui a menor prioridade. A associação de prioridades as mensagens que chegam ao sistema é feita pelo tipo da mensagem (voz, vídeo, dados, etc). Em cada estação, a chegada de mensagens é caracterizada por um processo de Poisson tal que, λ_i^p (mensagens por segmento) é a taxa média de chegada das mensagens de classe p na estação i . O número de pacotes que compõem a m -ésima mensagem de classe p é denotado por $B_{i,m}^p$ ($i = 1, 2, \dots, M; p = 1, 2, \dots, P$). Assume-se que $\{B_{i,m}^p; m \geq 1\}$ é uma seqüência de variáveis aleatórias independentes e identicamente distribuídas (i.i.d.) para cada p ; e $\beta_{i,j}^p = P(B_{i,m}^p = j)$ ($j = 1, 2, \dots$) é a distribuição de probabilidade de $B_{i,m}^p$, com média b_i^p e segundo momento $b_{2,i}^p$.

Como definido no Capítulo 5, $L_n = L_n^R + L_n^{DL} + L_n^T$ é o tamanho do n -ésimo ciclo onde L_n^R é o tamanho do n -ésimo intervalo de reserva, L_n^{DL} é o tamanho do sub-quadro DL pertencente ao n -ésimo ciclo e L_n^T é o tamanho do n -ésimo intervalo

de transmissão. Define-se que $L_n^R = M$ (onde M é o número de estações na rede), com duração de $M\tau$ segundos.

$N_{i,k}^p$ representa o número de mensagens de classe p chegando ao terminal i durante o k -ésimo segmento. O conjunto $\{N_{i,k}^p; k \geq 1\}$ ($i = 1, 2, \dots, M; p = 1, 2, \dots, P$) também é uma seqüência de variáveis aleatórias i.i.d., governada por uma distribuição de Poisson com média λ_i^p (mensagens/segmento), independente do processo de chegada de outras classes. Das definições acima têm-se que, $N_{i,k}$ é o número total de mensagens chegando ao terminal i durante o k -ésimo segmento e $\{N_{i,k}; k \geq 1\}$ é uma seqüência de variáveis aleatórias Poisson i.i.d., com média:

$$\lambda_i = \sum_{p=1}^P \lambda_i^p; \text{ para cada } i \in \{1, 2, \dots, M\}.$$

$W_{i,m}^p$ representa o tempo de espera (medido em segmentos) para a m -ésima mensagem de classe p chegando na estação i , ou seja, o número de segmentos desde a chegada da m -ésima mensagem de classe p até o segmento onde começa a transmissão desta mensagem. O número de mensagens de classe p armazenadas no *buffer* da estação i no começo do n -ésimo ciclo é representado por $Z_{i,n}^p$. $N_{i,k,n}^p$ é o número de mensagens de classe p chegando no terminal i durante o k -ésimo segmento do n -ésimo ciclo; e $B_{i,m,n}^p$ é o número de pacotes contidos na m -ésima mensagem de classe p transmitida pelo terminal i durante o n -ésimo ciclo.

A Cadeia de Markov vetorial $\underline{Y}_i^p = \{\underline{Y}_{i,n}^p; n \geq 1\}$, representa o processo de estado do canal onde $\underline{Y}_{i,n}^p = \{L_n, N_{i,k,n}^p, B_{i,m,n}^p\}$. Para análise do tempo de espera das mensagens define-se as seguintes funções de $\underline{Y}_{i,n}^p$:

- $N_i^p(\underline{Y}_{i,n}^p) =$ ao número total de mensagens de classe p transmitidas na estação i , durante o $(n + 1)$ -ésimo ciclo, dado $\underline{Y}_{i,n}^p$;
- $W_i^p(\underline{Y}_{i,n}^p) =$ a soma das componentes do tempo de espera de todas as mensagens de classe p servidas no terminal i no $(n + 1)$ -ésimo ciclo, dado $\underline{Y}_{i,n}^p$.

A partir das definições acima, têm-se o cálculo de $N_i^p(\underline{Y}_{i,n}^p)$ e $W_i^p(\underline{Y}_{i,n}^p)$ através das equações 6.1 e 6.2, respectivamente.

$$N_i^p(\underline{Y}_{i,n}^p) = Z_{i,n+1}^p; \quad i = 1, 2, \dots, M \text{ e } p = 1, 2, \dots, P \quad (6.1)$$

$$\begin{aligned} W_i^p(\underline{Y}_{i,n}^p) &= (N_{i,1,n}^p(L_n - 1) + N_{i,2,n}^p(L_n - 2) + \dots + N_{i,(L_n-1),n}^p) + \\ &+ (M + L_n^{DL})Z_{i,n+1}^p + \\ &+ \left(\sum_{j=1}^{p-1} \sum_{g=1}^M \sum_{k=1}^{Z_{g,n+1}^j} B_{g,k,n+1}^j + \sum_{g=1}^{i-1} \sum_{k=1}^{Z_{g,n+1}^p} B_{g,k,n+1}^p \right) Z_{i,n+1}^p + \\ &+ \left(B_{i,1,n+1}^p + (B_{i,1,n+1}^p + B_{i,2,n+1}^p) + \dots + \sum_{k=1}^{Z_{i,n+1}^p-1} B_{i,k,n+1}^p \right) \end{aligned} \quad (6.2)$$

O primeiro termo da equação (6.2) representa o atraso total das mensagens de classe p chegando ao terminal i durante L_n , desde os instantes de chegada até o término de L_n . O segundo termo é o atraso de todas as $Z_{i,n+1}^p$ mensagens devido aos intervalos de reserva e DL $(M + L_n^{DL})$. O terceiro termo é composto pelo tempo total que todas as $Z_{i,n+1}^p$ mensagens devem esperar devido às transmissões das mensagens de classe 1 até $(p - 1)$ em todos os terminais (1 até M), mais as transmissões das mensagens de classe p dos terminais 1 até $(i - 1)$. Por fim, o último termo representa o atraso total decorrido por todas as $Z_{i,n+1}^p$ mensagens entre os instantes em que a primeira destas mensagens começa a ser transmitida e início das demais transmissões (a primeira mensagem não espera nenhuma outra, a segunda espera pela transmissão da primeira, isto é $B_{i,n+1}^p$, e assim por diante).

Define-se o tempo médio de espera na fila para as mensagens de classe p na estação i como:

$$\bar{W}_i^p = \lim_{N \rightarrow \infty} N^{-1} \sum_{n=1}^N W_{i,n}^p.$$

Através das definições acima e utilizando a análise empregada em [19], têm-se a seguinte expressão para o tempo médio de espera na fila:

$$\bar{W}_i^p = \lim_{K \rightarrow \infty} \frac{\sum_{n=1}^K W_i^p(\underline{Y}_{i,n}^p)}{\sum_{n=1}^K N_i^p(\underline{Y}_{i,n}^p)}.$$

onde $\sum_{n=1}^K N_i^p(\underline{Y}_{i,n}^p) \rightarrow \infty$; com $K \rightarrow \infty$.

Através da aplicação do teorema *Markov Ratio Limit Theorem* - (MRLT) [48] na expressão anterior, obtém-se:

$$\bar{W}_i^p = \frac{E_o[W_i^p(\underline{Y}_{i,n}^p)]}{E_o[N_i^p(\underline{Y}_{i,n}^p)]}, \quad (6.3)$$

onde $E_o[N_i^p(\underline{Y}_{i,n}^p)]$ e $E_o[W_i^p(\underline{Y}_{i,n}^p)]$ são, respectivamente, as médias das equações (6.1) e (6.2) no estado de equilíbrio.

Através da equação (6.1) e da definição de $N_i^p(\underline{Y}_{i,n}^p)$, têm-se que o denominador da equação (6.3) é:

$$E_o[N_i^p(\underline{Y}_{i,n}^p)] = E_o[Z_{i,n+1}^p] = \lambda_i^p E_o[L_n] = \lambda_i^p E[L], \quad (6.4)$$

com $E[L] = \lim_{n \rightarrow \infty} E[L_n]$. Da mesma forma, o numerador da equação (6.3) é calculado aplicando-se a média termo a termo na equação (6.2). Com isso, o primeiro termo é dado por:

$$E_o[N_{i,1,n}^p(L_n - 1) + N_{i,2,n}^p(L_n - 2) + \dots + N_{i,(L_n-1),n}^p] = \frac{\lambda_i^p}{2}(E[L^2] - E[L]), \quad (6.5)$$

onde $E[L^2] = \lim_{n \rightarrow \infty} E[L_n^2]$. Aplicando a média no segundo termo da equação (6.2) temos,

$$\begin{aligned} E_o[(M + L_n^{DL})Z_{i,n+1}^p] &= (M + E_o[L_n^{DL}])\lambda_i^p E_o[L_n] \\ &= (M + E[DL])\lambda_i^p E[L], \end{aligned} \quad (6.6)$$

onde $E[DL] = \lim_{n \rightarrow \infty} E[L_n^{DL}]$. E, após algumas manipulações, têm-se as equações (6.7) e (6.8) referentes ao terceiro termo; e a equação (6.9) referente ao quarto termo:

$$\begin{aligned} E_o \left[\left(\sum_{j=1}^{p-1} \sum_{g=1}^M \sum_{k=1}^{Z_{g,n+1}^j} B_{g,k,n+1}^j \right) Z_{i,n+1}^p \right] &= \sum_{j=1}^{p-1} \sum_{g=1}^M b_g^j \lambda_g^j \lambda_i^p E[L^2] \\ &= \lambda_i^p E[L^2] \sum_{j=1}^{p-1} \sum_{g=1}^M \rho_g^j; \end{aligned} \quad (6.7)$$

$$E_o \left[\left(\sum_{g=1}^{i-1} \sum_{k=1}^{Z_{g,n+1}^j} B_{g,k,n+1}^p \right) Z_{i,n+1}^p \right] = \lambda_i^p E[L^2] \sum_{g=1}^{i-1} \rho_g^p \quad (6.8)$$

e

$$E_o \left[B_{i,1,n+1}^p + (B_{i,1,n+1}^p + B_{i,2,n+1}^p) + \dots + \sum_{k=1}^{Z_{i,n+1}^p - 1} B_{i,k,n+1}^p \right] = \frac{\rho_i^p \lambda_i^p E[L^2]}{2}, \quad (6.9)$$

onde $\rho_i^p = \lambda_i^p b_i^p$ nas equações (6.7), (6.8) e (6.9) é o tráfego no terminal i devido as mensagens de classe p . Das equações (6.5) à (6.9) obtêm-se $E_o[W_i^p(\underline{Y}_{i,n}^p)]$. Usando também $E_o[N_i^p(\underline{Y}_{i,n}^p)]$ dado pela equação (6.4), e substituindo os termos na equação (6.3), tem-se o tempo médio de espera para as mensagens de classe p na estação i para a Versão I do protocolo proposto:

$$\bar{W}_i^p = M + E[DL] + (1 + \rho_i^p) \frac{E[L^2]}{2E[L]} + \left(\sum_{j=1}^{p-1} \sum_{g=1}^M \rho_g^j + \sum_{j=1}^{i-1} \rho_j^p \right) \frac{E[L^2]}{E[L]} - \frac{1}{2}. \quad (6.10)$$

Nota-se que a expressão para \bar{W}_i^p é dada em função de $E[L]$ e $E[L^2]$, que são, respectivamente, o primeiro e o segundo momento do tamanho do ciclo de transmissão. Prosseguindo com a análise para encontrar uma fórmula fechada para o tempo médio de atraso na fila, tem-se que:

$$\begin{aligned}
E[L_n] &= E[L_n^R + L_n^{DL} + L_n^T] = E[L_n^R] + E[L_n^{DL}] + E[L_n^T] \\
&= M + E[L_n^{DL}] + E \left[\sum_{p=1}^P \sum_{i=1}^M \sum_{k=1}^{Z_{i,n}^p} B_{i,k,n}^p \right] \\
&= M + E[L_n^{DL}] + \sum_{p=1}^P \sum_{i=1}^M b_i^p E[Z_{i,n}^p],
\end{aligned}$$

lembrando que $E[Z_{i,n}^p] = \lambda_i^p E[L_{n-1}]$. Assim, obtém-se a seguinte expressão recursiva para $E[L_n]$:

$$E[L_n] = M + E[L_n^{DL}] + \sum_{p=1}^P \sum_{i=1}^M \rho_i^p E[L_{n-1}]. \quad (6.11)$$

Assumindo que $\rho = \sum_{p=1}^P \sum_{i=1}^M \rho_i^p < 1$ e tomando os limites (com $n \rightarrow \infty$) em ambos os lados da equação (6.11), obtém-se $E[L]$ pela equação 6.12.

$$E[L] = \frac{M + E[DL]}{1 - \rho}; \quad \rho = \sum_{p=1}^P \sum_{i=1}^M \rho_i^p < 1 \quad (6.12)$$

De maneira similar, através de uma equação recursiva em $E[L_n^2]$ e com $\rho < 1$, tem-se a seguinte expressão para $E[L^2]$:

$$\begin{aligned}
E[L^2] &= \frac{1}{1 - \sum_{p=1}^P \sum_{i=1}^M (\rho_i^p)^2} \left\{ M^2 + E[DL^2] + 2ME[DL] + \right. \\
&\quad + \left[2(M + E[DL]) \sum_{p=1}^P \sum_{i=1}^M \rho_i^p + \sum_{p=1}^P \sum_{i=1}^M \lambda_i^p b_{2,i}^p \right] E[L] + \\
&\quad \left. + \left[\sum_{p=1}^P \sum_{i=1}^M \sum_{\substack{j=1 \\ j \neq i}}^M \rho_i^p \rho_j^p + \sum_{p=1}^P \sum_{\substack{q=1 \\ q \neq p}}^P \sum_{i=1}^M \sum_{k=1}^M \rho_i^p \rho_k^q \right] E[L]^2 \right\}. \quad (6.13)
\end{aligned}$$

Finalmente, substituindo as equações (6.12) e (6.13) na equação (6.10), tem-se uma expressão fechada para o tempo médio de espera das mensagens de classe p na estação i com a Versão I do protocolo proposto.

6.2 Modelagem da Versão II

A análise para a Versão II do protocolo proposto segue de maneira direta notando-se que, de acordo com este esquema, as mensagens são transmitidas na mesma ordem que a Versão I, com as classes das mensagens trocadas pelos números das estações e vice-versa (vide Figuras 5.3 e 5.4). Então, a expressão para \overline{W}_i^p na Versão II é análoga a da Versão I, alterando-se apenas o i pelo p e o M pelo P e vice-versa. Assim, obtemos a seguinte expressão para o tempo médio de espera das mensagens de classe p na estação i com a Versão II do protocolo proposto:

$$\overline{W}_i^p = M + E[DL] + (1 + \rho_i^p) \frac{E[L^2]}{2E[L]} + \left(\sum_{j=1}^{i-1} \sum_{k=1}^P \rho_j^k + \sum_{j=1}^{p-1} \rho_i^j \right) \frac{E[L^2]}{E[L]} - \frac{1}{2}; \quad (6.14)$$

com $E[L]$ e $E[L^2]$ dados pelas equações (6.12) e (6.13), respectivamente.

Porém, como será visto no próximo capítulo, a Versão II do protocolo proposto apresenta uma maior “injustiça” no acesso ao meio entre as estações da rede. Isto se deve à precedência estática entre as estações, ou seja, ao esquema de prioridade fixa que determina a ordem de transmissão durante o sub-canal UL. Como a Versão II prioriza primeiro as estações e não os fluxos, o terminal de menor prioridade só transmite depois de todos os fluxos dos outros terminais.

Para solucionar este problema, foi incorporado a Versão II do protocolo proposto duas disciplinas de prioridades variáveis entre as estações da rede: uma com prioridade cíclica e outra com prioridade complementar. Estes esquemas baseiam-se no escalonamento proposto em [32], porém, incluindo-se as prioridades por classes de tráfego.

Prioridade Cíclica na Versão II

Esta disciplina define que, uma estação que tem prioridade p ($p \in 1, 2, \dots, M$) em um determinado quadro, assume a prioridade $p + 1$ (uma menor) no próximo quadro. A estação com menor prioridade (M) obtém a maior prioridade no próximo

quadro.

Denota-se $\overline{W}_i^p(x)$ como o tempo médio de espera na fila para as mensagens de classe p na estação i , dado que a estação possui prioridade x . Além disso, assume-se que $F_i^p(x)$ é a probabilidade que uma mensagem de classe p na estação i seja transmitida durante o quadro em que a estação i possui prioridade x . Dessa forma, têm-se que:

$$\overline{W}_i^p = \sum_{x=1}^M \overline{W}_i^p(x) \cdot F_i^p(x). \quad (6.15)$$

$F_i^p(x)$ é a probabilidade que a mensagem chegue durante o quadro em que a estação i possui prioridade $x - 1$ (ou prioridade M , caso $x = 1$). Como os quadros são identicamente distribuídos (independente da disciplina de prioridade utilizada) e supõem-se chegadas de Poisson, conclui-se que [32]:

$$F_i^p(1) = F_i^p(2) = \dots = F_i^p(M) = \frac{1}{M}. \quad (6.16)$$

Utilizando o resultado obtido na equação 6.14, tem-se $\overline{W}_i^p(x)$ através da equação 6.17:

$$\begin{aligned} \overline{W}_i^p(x) = & M + E[DL] + (1 + \rho_i^p) \frac{E[L^2]}{2E[L]} + \\ & + \left(\sum_{j=1}^{x-1} \sum_{k=1}^P \rho_{1+(i-x+j-1) \bmod(M)}^k + \sum_{j=1}^{p-1} \rho_i^j \right) \frac{E[L^2]}{E[L]} - \frac{1}{2}. \end{aligned} \quad (6.17)$$

Substituindo as equações 6.16 e 6.17 na equação 6.15, tem-se o tempo médio de espera na fila para as mensagens de classe p na estação i , sob a disciplina de prioridade cíclica (equação 6.18):

$$\begin{aligned} \bar{W}_i^p &= M + E[DL] + (1 + \rho_i^p) \frac{E[L^2]}{2E[L]} + \\ &+ \left(\frac{1}{M} \sum_{j=2}^M (M - j + 1) \sum_{k=1}^P \rho_{1+(i-j) \bmod(M)}^k + \sum_{j=1}^{p-1} \rho_i^j \right) \frac{E[L^2]}{E[L]} - \frac{1}{2}. \end{aligned} \quad (6.18)$$

Prioridade Complementar na Versão II

Esta disciplina define que, a estação i assume prioridade x em um quadro e prioridade $(M + 1 - x)$ no próximo; depois retorna para prioridade x e assim sucessivamente. Utilizando a mesma análise empregada para a disciplina cíclica pode-se mostrar que, através da disciplina complementar o tempo médio de espera na fila é o mesmo para todas as estações, de acordo com a equação 6.19:

$$\bar{W}_i^p = M + E[DL] + (1 + \rho_i^p) \frac{E[L^2]}{2E[L]} + \left(\frac{1}{2} \sum_{\substack{j=1 \\ j \neq i}}^M \sum_{q=1}^P \rho_j^q + \sum_{k=1}^{p-1} \rho_i^k \right) \frac{E[L^2]}{E[L]} - \frac{1}{2}. \quad (6.19)$$

Através da modelagem analítica apresentada neste capítulo, alguns resultados numéricos serão apresentados em seguida, para ilustrar o desempenho dos protocolos propostos com prioridades baseadas em mensagem e/ou estações.

Capítulo 7

Resultados Obtidos

PARA analisar o comportamento do protocolo proposto com relação a diferenciação do serviço oferecido para as diferentes classes de tráfego, este capítulo apresenta alguns resultados numéricos obtidos com os modelos analíticos descritos no capítulo anterior. Além disso, através de uma ferramenta de simulação, foi desenvolvido um modelo de fila para validar a solução proposta. Como esperado, observa-se que o método proposto consegue diferenciar o serviço oferecido para as classes de tráfego, aplicando as regras de prioridade estabelecidas.

7.1 Cenários de Avaliação

Para avaliar o nível de diferenciação obtido com os protocolos descritos serão considerados dois cenários distintos, onde em cada cenário existe uma probabilidade diferenciada entre quatro tipos de classes de tráfego ($P = 4$), como mostra a Tabela 7.1. Estas quatro classes podem ser mapeadas para os quatro tipos de serviços oferecidos pelo padrão IEEE 802.16 da seguinte maneira: a classe 1 (que é a mais prioritária) representa os serviços UGS, a classe 2 representa os serviços rtPS, a classe 3 representa os serviços nrtPS e a classe 4 representa os serviços BE. A diferença entre os cenários é que, no Cenário I existe uma maior probabilidade para as classes de maior prioridade, enquanto que, no Cenário II as classes de menor prioridade predominam sobre as de maior prioridade. Assim, pode-se comparar qual é a influência de uma alta carga dos fluxos de menor prioridade sobre os de maior prioridade e vice-versa.

Classe de Tráfego	Cenário I	Cenário II
Classe 1	40%	10%
Classe 2	30%	20%
Classe 3	20%	30%
Classe 4	10%	40%

Tabela 7.1: Cenários de tráfego utilizados na modelagem analítica.

Em cada cenário existem 10 estações ($M = 10$) com o tráfego balanceado entre as mesmas, ou seja, $\lambda_i^p = \lambda^p/10$, onde λ^p representa a taxa de mensagens da classe p e λ_i^p a taxa de mensagens da classe p na estação i . Assume-se que o número de pacotes em cada mensagem de classe p na estação i é geometricamente distribuído com média $b_i^p = 5$ e $b_{2,i}^p = 45$, para cada $p = 1, 2, 3, 4$; e $i = 1, \dots, 10$. Com isso, o tempo médio de espera para a classe p é dado por:

$$\bar{W}^p = \sum_{i=1}^{10} \frac{\lambda_i^p}{\lambda^p} \bar{W}_i^p.$$

As Figuras 7.1 e 7.2 ilustram o tempo médio de espera na fila para cada classe de

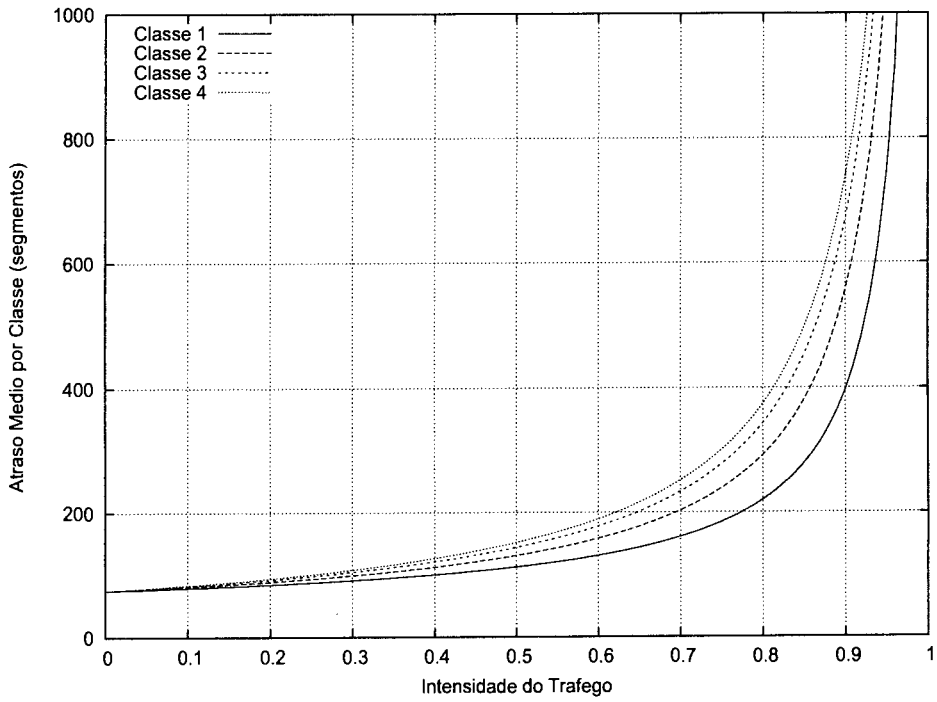
prioridade em relação ao tráfego oferecido no canal. O comportamento das Versões I e II do protocolo proposto sob o Cenário I é apresentado nas Figuras 7.1(a) e (b) respectivamente. Da mesma forma, as Figuras 7.2(a) e (b) mostram o comportamento das duas versões sob o Cenário II.

Pelas figuras, observa-se uma diferenciação mais evidente para um alto tráfego no canal e que, com o aumento da intensidade do tráfego, cresce o tempo de espera na fila para todas as classes. Porém, esta diferenciação é menor para Versão II do protocolo proposto, como mostram as Figuras 7.1(b) e 7.2(b). Isto acontece porque, na Versão I as prioridades entre classes predominam sobre as prioridades entre estações, ocorrendo o contrário na Versão II onde as prioridades entre estações sobrepõem-se. Através das Figuras 7.1(a) e 7.2(a) observa-se que, o tempo de espera para o tráfego de alta prioridade (classe 1) é menor em relação as demais classes, até mesmo no Cenário II onde existe uma maior probabilidade dos tráfegos de baixa prioridade. Com isso, pode-se perceber que os protocolos propostos conseguem diferenciar eficientemente as classes de tráfego, garantindo menor tempo de espera na fila para as mensagens de maior prioridade.

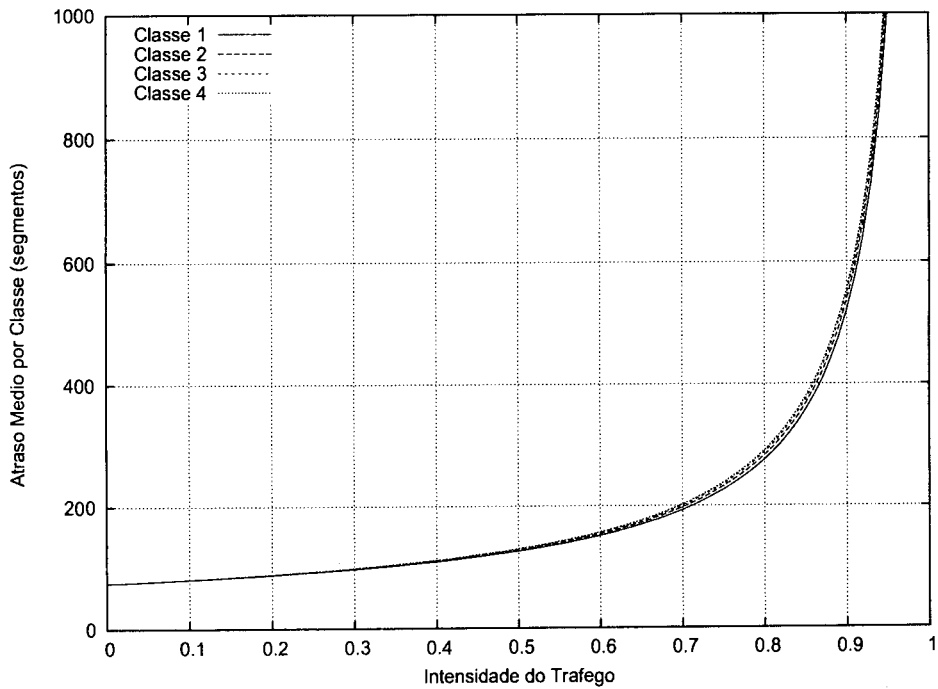
As Figuras 7.3 e 7.4 apresentam o tempo médio de espera na fila em função dos terminais, para três valores de intensidade de tráfego ($\rho = 0,3; 0,6$ e $0,9$). Dessa forma, o tempo médio de espera na fila para estação i é dado pela equação abaixo, onde λ_i representa a taxa de mensagens na estação i .

$$\bar{W}_i = \sum_{p=1}^4 \frac{\lambda_i^p}{\lambda_i} \bar{W}_i^p$$

O comportamento das duas versões do protocolo proposto no Cenário I está ilustrado nas Figuras 7.3(a) e (b). Em seguida, as Figuras 7.4(a) e (b) apresentam, respectivamente, as Versões I e II com a utilização do Cenário II. Das Figuras 7.3(a) e (b), observa-se que, para valores fixos de p , a diferença do tempo de espera na fila entre as estações na Versão I é menor do que na Versão II pois, como já foi dito anteriormente, na Versão I a priorização é feita primeiro por classes e não por estações. Por isso, a variação do tempo de espera entre classes é menor para a Versão

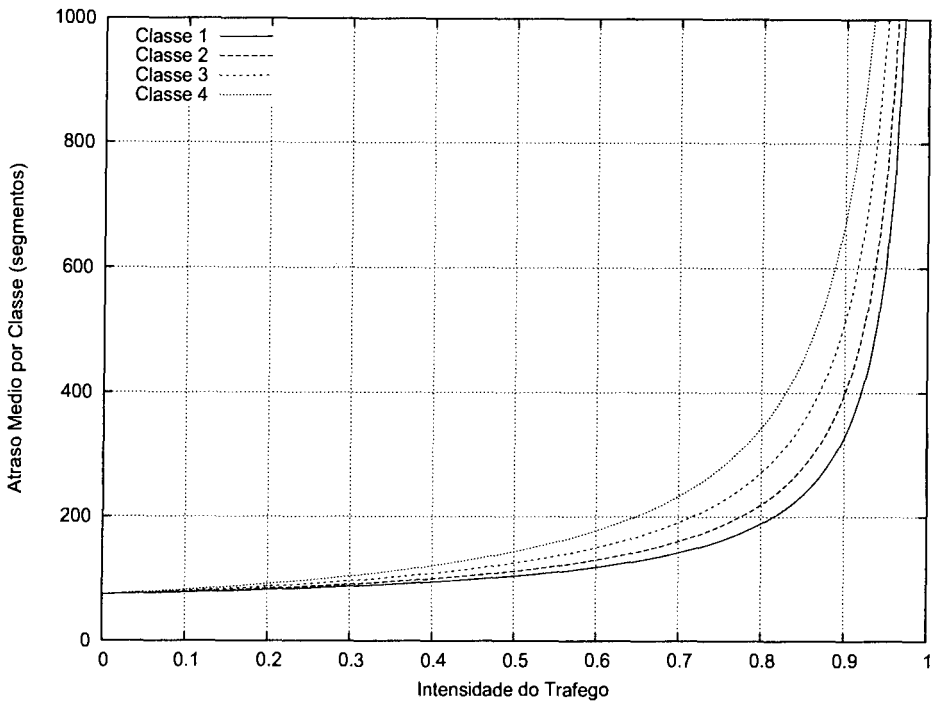


(a) Versão I

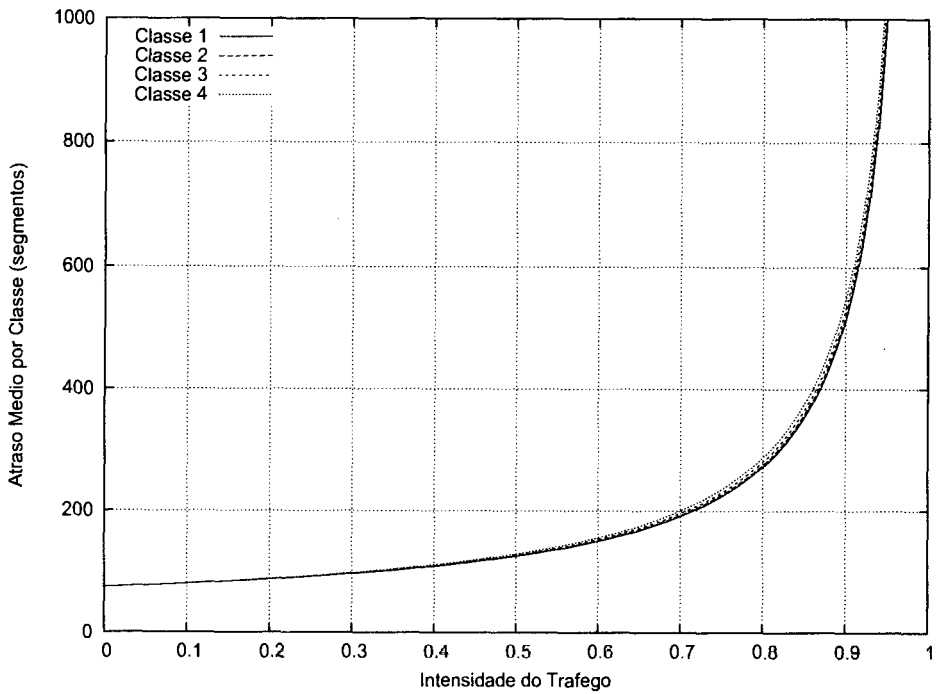


(b) Versão II

Figura 7.1: Tempo médio de espera na fila para cada classe de prioridade versus tráfego oferecido no Cenário I: Versão I (a) e Versão II (b).

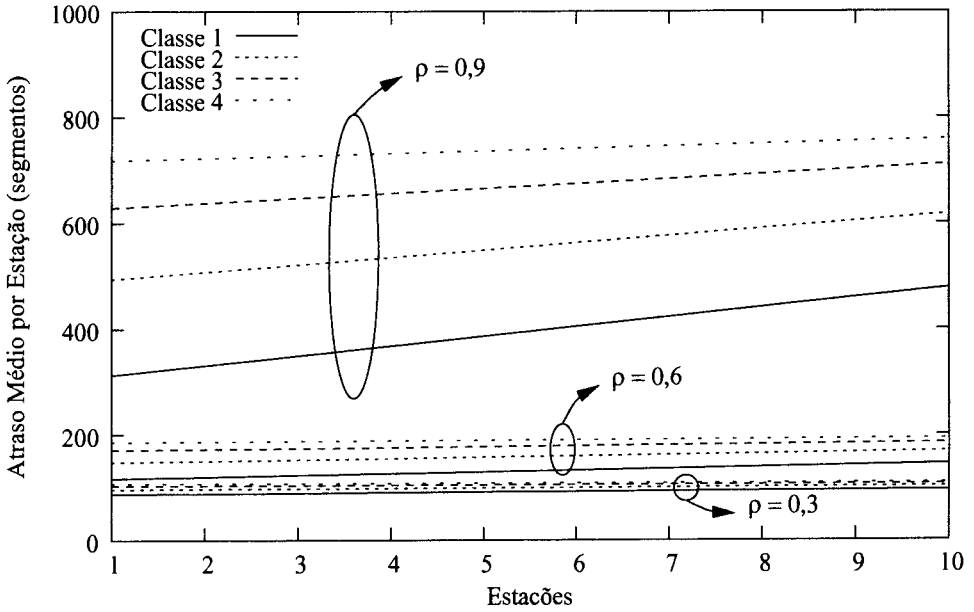


(a) Versão I

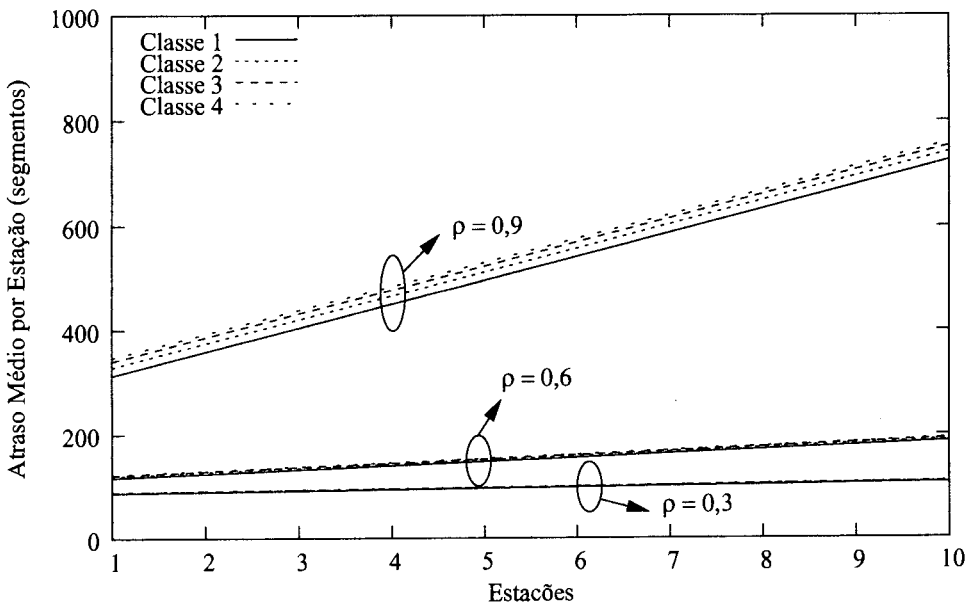


(b) Versão II

Figura 7.2: Tempo médio de espera na fila para cada classe de prioridade versus tráfego oferecido no Cenário II: Versão I (a) e Versão II (b).



(a) Versão I



(b) Versão II

Figura 7.3: Tempo médio de espera na fila para diferentes valores de ρ em cada estação no Cenário I: Versão I (a) e Versão II (b).

