COPPE/UFRJ

MODELAGEM ANALÍTICA E AVALIAÇÃO DO RETARDO DAS MENSAGENS NO PROTOCOLO DE ACESSO AO MEIO DO PADRÃO IEEE 802.16

Danielle Lopes Ferreira Gonçalves Vieira

Dissertação de Mestrado apresentada ao Programa de Pós-Graduação em Engenharia de Sistemas e Computação, COPPE, da Universidade Federal do Rio de Janeiro, como parte dos requisitos necessários à obtenção do título de Mestre em Engenharia de Sistemas e Computação.

Orientador(es): Luís Felipe Magalhães de Moraes

Rio de Janeiro

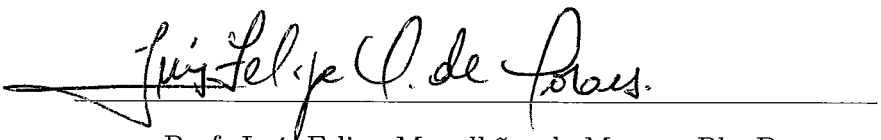
Setembro de 2008

MODELAGEM ANALÍTICA E AVALIAÇÃO DO RETARDO DAS MENSAGENS NO PROTOCOLO DE ACESSO AO MEIO DO PADRÃO IEEE 802.16

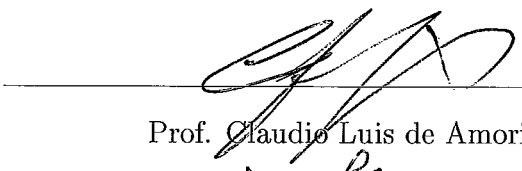
Danielle Lopes Ferreira Gonçalves Vieira

DISSERTAÇÃO SUBMETIDA AO CORPO DOCENTE DO INSTITUTO ALBERTO LUIZ COIMBRA DE PÓS-GRADUAÇÃO E PESQUISA 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.

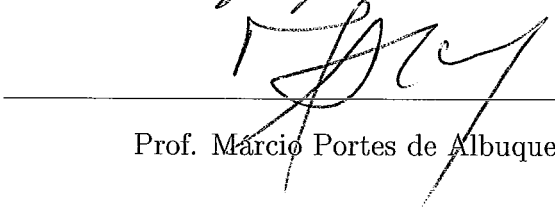
Aprovada por:



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



Prof. Claudio Luis de Amorim, Ph. D.



Prof. Márcio Portes de Albuquerque, Dr.

RIO DE JANEIRO, RJ - BRASIL

SETEMBRO DE 2008

Vieira, Danielle Lopes Ferreira Gonçalves

Modelagem Analítica e Avaliação do Retardo das Mensagens no Protocolo de Acesso ao Meio do Padrão IEEE 802.16/Danielle Lopes Ferreira Gonçalves Vieira. - Rio de Janeiro: UFRJ/COPPE, 2008.

XVIII, 93 p.: il.; 29,7 cm.

Orientador: Luís Felipe Magalhães de Moraes

Dissertação (mestrado) - UFRJ/COPPE/Programa de Engenharia de Sistemas e Computação, 2008.

Referências Bibliográficas: p. 84-88.

1. Redes Metropolitanas Sem Fio. 2. Protocolos de Acesso ao Meio. 3. Qualidade de Serviço. 4. Avaliação de Desempenho I. Moraes, Luís Felipe Magalhães de II. Universidade Federal do Rio de Janeiro, COPPE, Programa de Engenharia de Sistemas e Computação. III. Título.

Dedicatória

Dedico esse trabalho as minhas lindas e admiráveis filhas Eduarda e Pietra, que são a razão de minha vida. Desculpem as horas roubadas de nosso convívio, pelo tempo dedicado ao estudo, concedendo a mim a oportunidade de me realizar ainda mais.

Agradecimentos

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

Ao meu pai Jorge, pelas palavras sábias e incentivadoras nos momentos difíceis, que me deram força e coragem para continuar. Agradecimento mais que especial a minha mãe, pelo apoio, suporte, amor e preocupação, sempre. Sem você eu não teria conseguido. Carinhosamente ao meu marido Edward. Aos meus irmãos Jefferson, Vinícius e Débora (em especial a querida irmã e amiga Débora, pelo apoio incondicional ao meu estudo), as minhas filhas Eduarda e Pietra, ao cunhado Serginho, bem como, toda a minha família, por todo amor e carinho, não só durante a realização deste trabalho, como na vida inteira.

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 Claudio Amorim e Márcio Portes, pela valiosa ajuda nesta fase final.

Agradeço a todos os amigos que conheci durante este período, em particular: Tiago, Rafael Fernandes, Rafael Bezerra, Júlio, Paulo, Bruno, Eduardo, Airon, Cláudia, Diogo, Verissimo, Schiller, Michelini e todos os outros, que por ventura eu tenha esquecido. Agradecimento especial ao amigo Gustavo, pela ajuda, contribuição, companheirismo e amizade. Agradecimento muito especial ao amigo Beto, pela amizade, pela ajuda e pelas grandes contribuições na reta final de elaboração dessa dissertação.

Ao Programa de Engenharia de Sistemas e Computação (PESC/COPPE/UFRJ), pelo apoio operacional.

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

MODELAGEM ANALÍTICA E AVALIAÇÃO DO RETARDO DAS MENSAGENS NO PROTOCOLO DE ACESSO AO MEIO DO PADRÃO IEEE 802.16

Danielle Lopes Ferreira

Setembro/2008

Orientador: Luís Felipe Magalhães de Moraes

Programa: Engenharia de Sistemas e Computação

O padrão IEEE 802.16 surgiu como uma solução para o acesso à banda larga através de meios sem fio, sendo desenvolvido para transmitir dados e aplicações multimídia com diferentes requisitos de qualidade de serviço (QoS). Visando alcançar o nível de QoS desejado para aplicações multimídia, esse padrão fornece diferentes mecanismos de escalonamento. Vários trabalhos na literatura investigam o impacto dos mecanismos de escalonamento no desempenho dessas redes. A maioria desses trabalhos apresentam resultados de avaliações de desempenho obtidos através de simulação ou propõem modelos analíticos para essa avaliação, mas com alguma alteração do protocolo de acesso ao meio (MAC) do padrão IEEE 802.16. Dentro desse contexto, este trabalho modela o retardo total das mensagens de aplicações em tempo real e outros, transmitidas pelo protocolo MAC do padrão IEEE 802.16. O retardo total das mensagens é obtido através de dois modelos distintos: um para a fase de contenção (disputa pelo meio físico) aplicado somente ao tráfego de tempo não-real; e outro para a fase de alocação de dados, onde o modelo suporta duas classes de prioridade de tráfego, uma para serviço de tempo real e outra para tempo não-real. Além disso, foi realizada uma avaliação da solução proposta, onde foi analisado o retardo total das mensagens para dois cenários, na qual a carga de cada tipo varia. Também foi investigado a influência de alguns parâmetros, do mecanismo de acesso aleatório, no desempenho destas redes. Após a avaliação da solução proposta, os resultados obtidos com o modelo analítico são comparados com resultados obtidos através de simulação. Onde observou-se que o modelo analítico proposto representa o comportamento do protocolo da camada MAC do padrão IEEE 802.16, considerando como métrica de desempenho o retardo total do sistema.

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

ANALYTICAL MODELLING AND EVALUATION OF THE DELAY OF THE MESSAGES IN THE PROTOCOL OF ACCESS TO IEEE 802.16 STANDARD

Danielle Lopes Ferreira

September/2008

Advisor: Luís Felipe Magalhães de Moraes

Department: Systems Engineering and Computer Science

The IEEE 802.16 standard came up as a solution to broadband access through wireless media, being developed to transmit data and multimedia applications with different quality of service (QoS) requirements. In order to achieve the QoS requirements of multimedia applications, the IEEE 802.16 standard offers different scheduling schemes. Thus, lots of work found in literature investigates the impact of scheduling mechanisms on the performance of these networks, however these works present the performance evaluations through simulation. Moreover, some works propose analytical models to perform this evaluation, but with some change in the IEEE 802.16 standard media access protocol (MAC). Thus, this work models the total delay of messages for the real and non-real time traffic transmitted by the IEEE 802.16 standard. The total delay of messages is obtained through two distinct models: one for contention phase (disputation by physical means), applied only for the non-real time traffic; and another model for the data allocation phase, where the model support two classes of traffic priority, one for real time service and another for non-real time. Moreover, an evaluation of the presented solution was accomplished, where the total delay of messages was analyzed for two scenaries, which the load of each type varies. Besides, the influence of some parameters was investigated, of the random access mechanism, in performance of the networks. After an evaluation of the presented solution, the results obtained with the analytical model are compared with the simulation results. Where it was observed that the proposed analytical model represents the behavior of the IEEE 802.16 standard MAC layer protocol, as a function of the systems end-to-end delay.

Conteúdo

Lista de Figuras

Lista de Tabelas

Lista de Acrônimos

ATM	:	*Asynchronous Transfer Mode*;
BS	:	*Base Station*;
DL	:	*Downlink*;
DSL	:	*Digital Subscriber Line*;
ETSI	:	*European Telecommunications Standards Institute*;
FDD	:	*Frequency Division Duplex*;
FIFO	:	*First In, First Out*;
FCFS	:	*First Come, First Serverd*;
HOL	:	*Head of the Line*;
IE	:	*Information Element*;
IEEE	:	*Institute of Electrical and Electronic Engineers*;
MAC	:	*Medium Access Control*;
PHY	:	*Physical Layer*;
PMP	:	*Point Multi-Point*;
QoS	:	*Quality of Service*;
SS	:	*Subscriber Station*;
TDD	:	*Time-Division Duplex*;
TDM	:	*Time-Division Multiplexing*;
TDMA	:	*Time-Division Multiple Access*;
UL	:	*Uplink*;
WLAN	:	*Wireless Local Area Network*;
WMAN	:	*Wireless Metropolitan Area Network*;
WPAN	:	*Wireless Personal Area Network*;

Notações do Modelo Analítico

N	:	*Número de estações no sistema*;
BE	:	*Expoente de backoff*;
N_B	:	*Número de tentativas de backoff*;
m	:	*estágio máximo de backoff*;
R	:	*Número de tentativas no estágio máximo de backoff*;
$B(t)$	:	*Processo estocástico que representa o tamanho do contador de tempo de backoff para uma determinada estação*;
$S(t)$	:	*Processo estocástico que representa o estágio de backoff no tempo t*;
τ	:	*Probabilidade estacionária de uma estação transmitir uma mensagem num segmento de tempo aleatório*;
p	:	*Probabilidade de colisão de uma mensagem de requisição de largura de banda*;
P_{tr}	:	*Probabilidade de pelo menos uma estação transmitir durante um segmento escolhido aleatoriamente*;
P_s	:	*Probabilidade de transmissão com sucesso*;
W_{min}	:	*Janela mínima de backoff*;
W_{max}	:	*Janela máxima de backoff*;
i	:	*Estágio de backoff*;
k	:	*Contador de backoff - número de oportunidades de transmissão para o qual uma estação deve esperar antes de iniciar a sua transmissão*;

$\{S(t), B(t)\}$	:	*Processo bi-dimensional representando a cadeia de Markov de tempo discreto do algoritmo de backoff*;
$b_{i,k}$	:	*Distribuição estacionária das probabilidades de transição da cadeia de Markov bi-dimensional*;
D_r	:	*Variável aleatória que representa o retardo médio das mensagens de requisição de largura de banda*;
D_1	:	*Variável aleatória que representa o retardo médio para as mensagens de tempo real*;
D_2	:	*Variável aleatória que representa o retardo médio para as mensagens de tempo não-real*;
N_c	:	*Variável aleatória que representa o número médio de colisões de uma mensagem de requisição de largura de banda*;
T_c	:	*Tempo de duração de uma colisão da mensagem de requisição de largura de banda*;
T_s	:	*Tempo gasto para uma transmissão com sucesso*;
β	:	*Variável aleatória que representa o retardo médio do contador de backoff*;
Φ	:	*Variável aleatória que representa o tempo pelo qual o contador da estação permanece congelado devido a transmissão de outras estações*;
N_q	:	*Variável aleatória que representa o número de vezes que uma estação deve esperar pela oportunidade de transmissão de outras estações antes do seu contador alcançar 0*;
μ	:	*Tamanho do período de contenção*;
λ_1	:	*Taxa de chegada das requisições virtuais de tempo real*;
λ_2	:	*Taxa de chegada das requisições de tempo não-real*;
ν_1	:	*Tempo de serviço médio das requisições virtuais de tempo real*;
ν_2	:	*Tempo de serviço médio das requisições de tempo não-real*;
l	:	*Número de requisições para o tráfego de tempo não-real*;

g : *Número de requisições de tempo real que chegam antes da i-ésima requisição de tempo não-real ser servida*;

S_i : *Variável aleatória que representa o tempo de serviço para a i-ésima requisição para o tráfego de tempo não-real*;

V_j : *Variável aleatória que representa o tempo de serviço para a j-ésima requisição virtual para o tráfego de tempo real*;

t_2 : *Tempo do início do próximo MAP*;

t' : *Tempo de chegada da i-ésima requisição para o tráfego de tempo não-real*;

T_{MAP} : *variável aleatória que representa o tempo médio definido pelo MAP*;

L : *Variável aleatória que representa o número de requisições virtuais para o tráfego de tempo real chegando durante o tempo do próximo MAP*;

G : *Variável aleatória que representa o número de requisições para o tráfego de tempo não-real chegando de* t_1 *até* t_2;

ρ_1 : *Carga oferecida do tráfego de tempo real*;

ρ_2 : *Carga oferecida do tráfego de tempo não-real*;

ρ : *Carga total oferecida*;

δ : *Variável aleatória que representa o retardo do contador de backoff de uma estação antes de acessar o canal em condições ocupadas*;

X : *Variável aleatória*;

$E[X]$: *Média da variável aleatória X*

$E[X^2]$: *Segundo momento da variável aleatória X*

Capítulo 1

Introdução

A RÁPIDA proliferação dos dispositivos móveis (como *laptops*, *handhelds* e PDAs) conduziu a uma mudança revolucionária na área da computação nos últimos anos. As redes sem fio formam atualmente uma grande vertente tecnológica, justificada pela busca de praticidade e acessibilidade aos meios de comunicação. Este capítulo introdutório apresenta os conceitos básicos inerentes às comunicações sem fio. A primeira seção apresenta uma visão geral de redes sem fio através das padronizações existentes para esta tecnologia. A motivação e os objetivos do trabalho são expostos nas próximas seções. Além disso, as principais contribuições alcançadas são apresentadas. Por fim, a estrutura do trabalho está descrita na última seção.

1.1 Redes sem Fio

A implantação da infraestrutura necessária à comunicação sem fio através da instalação de pontos de acesso à rede sem fio, nos quais pode-se acessar a Internet, nos principais grandes centros, são uma grande tendência e vêm apresentando um crescimento extremamente rápido. O elevado aumento no número de dispositivos de computação móveis, como *laptops*, *handhelds* e PDAs, conduziu a uma mudança revolucionária na computação nos últimos anos. A era do computador pessoal (um computador por pessoa) está perdendo espaço para a era da "computação úbiqua", na qual usuários utilizam, simultaneamente, vários aparelhos eletrônicos através dos quais podem acessar todas as informações necessárias a qualquer hora e em qualquer lugar, fazendo com que a comunicação, através de redes sem fio, seja a solução mais simples para os interconectar.

Uma rede sem fio pode oferecer conexão aos serviços fornecidos na Internet além de permitir a conectividade sem fio de estações fixas, móveis e portáteis, dentro de uma determinada região.

Classificação, Padrões e Tecnologias das Redes sem fio

- WPAN - *Wireless Personal Area Network* ou rede pessoal sem fio. Abrange o ambiente que cerca o usuário, com um alcance de aproximadamente 10 metros. É em geral utilizada para interligar dispositivos eletrônicos fisicamente próximos, eliminando os cabos usualmente utilizados para interligar teclados, impressoras, telefones móveis, agendas eletrônicas, computadores de mão, mouses e outros. Alcança taxas de alguns kbps até alguns Mbps. Como exemplo de tecnologia pode-se citar o *bluetooth*, padrão IEEE 802.15 [1], para estabelecer esta comunicação.

- WLAN - *Wireless Local Area Network* ou rede local sem fio. É utilizada para interligar dispositivos sem fio, cujo alcance de transmissão chega a algumas centenas de metros. As taxas de transmissão são da ordem de dezenas de Mbps. 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, também conhecido com WiFi, é mais difundido e faz parte da aliança internacional de fabricantes WECA (*Wireless Ethernet Compatibility Alliance*).

- WMAN - *Wireless Metropolitan Area Network* ou redes metropolitanas sem fio. Tem o alcance maior que as WLANs, chegando a dezenas de quilômetros e taxas de transmissão da ordem de centenas de Mbps. Há duas padronizações para as WMANs: o padrão IEEE 802.16 [4] e o padrão ETSI HIPERMAN [5]. O padrão IEEE 802.16 faz parte de uma aliança internacional denominada WiMax [6], tendo como objetivo garantir a interoperabilidade entre os dispositivos de diferentes fabricantes.

Dentre as redes sem fio, a crescente demanda por acesso à Internet com alta velocidade e serviços de multimídia, 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 via cabo e DSL (*Digital Subscriber Line*), dentre as quais: rápida implantação, alta escalabilidade, baixo custo de atualização e manutenção. O padrão IEEE 802.16 [4], surgiu 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).

Para fornecer essa qualidade de serviço, o protocolo de controle de acesso ao meio (MAC) do padrão IEEE 802.16 oferece diferentes formas de compartilhamento do meio sem fio, para os diferente tipos de tráfego. Porém, sabe-se que um dos grandes problemas que surgem nas redes de comunicação 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 forma que a faixa de freqüência da transmissão seja dividida eficientemente entre os usuários. A solução mais adequada depende das características do ambiente em questão e dos requisitos a serem atendidos.

A parte mais crítica da camada MAC de uma tecnologia sem fio infraestruturada é o múltiplo acesso no canal de subida (*uplink*). No canal de descida (*downlink*), a estação base tem completo conhecimento da demanda de largura de banda corrente, isto é, das mensagens armazenados no seu *buffer* e está apto a escalonar a transmissão. No *uplink*, a estação base não conhece o conteúdo do *buffer* das estações. Há duas soluções extremas para o múltiplo acesso no *uplink*. A primeira é alocar recursos para cada estação, tendo ou não dados a enviar. A outra possibilidade é garantir recursos para a estação apenas quando ela tiver dados prontos para enviar e requisitar largura de banda explicitamente. A vantagem do primeiro extremo é o pequeno retardo de acesso, a desvantagem é o desperdício do recurso se não houver dados a transmitir na estação. A desvantagem do segundo extremo é o retardo de acesso adicional e o recurso adicional para transmitir requisição de largura de banda, a vantagem é a boa utilização dos recursos. Como será apresentado no próximo capítulo, o protocolo da camada MAC do padrão IEEE 802.16 faz o uso destas duas soluções extremas para alocar recursos para os diferentes tipos de tráfego.

Portanto, a avaliação de desempenho do protocolo da camada MAC do padrão IEEE 802.16 torna-se de extrema importância e grande utilidade. Após a contextualização do presente trabalho serão descritas, nas seções abaixo, a motivação e definição do problema, os objetivos e as contribuições deste trabalho.

1.2 Motivação

A busca de soluções para a avaliação de desempenho das redes do padrão IEEE 802.16, é alvo de esforços da comunidade científica, como identificado na literatura. Diferentes técnicas, metodologias e teorias são conhecidas para avaliação de desempenho de sistemas computacionais. Cada uma apresenta possibilidades, vantagens e limitações, e são aplicáveis em diferentes contextos e com custo de avaliação diverso. A seguir, essas técnicas serão descritas e contextualizadas.

1. Medição - é uma técnica de medição de desempenho de sistemas reais e consiste em monitorar o sistema enquanto ele está sendo submetido a uma carga

em particular. Esta técnica é aplicada em um sistema que está em um estágio de desenvolvimento pós-protótipo. A ferramenta desta técnica é a instrumentação do sistema e apresenta uma precisão variável, além de um alto custo de implementação [7]. Um monitor é uma ferramenta utilizada para observar as atividades de um sistema. Em geral, os monitores coletam resultados de desempenho, produzem estatísticas e apresentam os resultados, além de identificarem áreas com problemas e sugerirem correções. Monitores são utilizados não apenas por analistas de desempenho, mas também por programadores e gerentes de sistemas. Monitoramento é o primeiro passo em medições de desempenho. A técnica de experimentação tem grande importância prática por identificar problemas correntes, como a necessidade de ajustes de parâmetros, além de prever potenciais problemas futuros. A principal vantagem é obter o desempenho do sistema real ao invéz do desempenho do modelo do sistema, pois as interações que afetam o desempenho do sistema real podem ser difíceis de se captar no modelo do sistema. Como desvantagem pode-se citar a necessidade de ter um sistema em execução e de instrumentar o sistema. Além disso, é difícil estimar o tempo gasto para instrumentar, realizar as medidas e modificar o sistema para estudar o efeito das alterações.

2. Modelos de simulação - a simulação pode ser utilizada para avaliar e modelar o desempenho de um sistema computacional e é uma técnica aplicada em qualquer estágio do ciclo de vida de um sistema, o tempo exigido para sua aplicação é médio. Esta técnica utiliza, como ferramenta, linguagens de programação e apresenta uma precisão moderada e um médio custo de implementação [7]. Um simulador é construído a partir de um modelo de desempenho do sistema. Entretanto, modelos de simulação podem falhar e muito tempo pode ser gasto em seu desenvolvimento. Escolher a linguagem é um importante passo no processo de desenvolvimento de um modelo de simulação [7]. Existem quatro opções: linguagens de simulação, linguagens de propósito geral, extensões de linguagens de propósito geral e pacotes de simulação [8]. Cada uma delas apresenta vantagens e desvantagens em relação ao consumo de tempo, facilidades embutidas na linguagem, até a familiaridade do progra-

mador e do analista com a linguagem. A simulação é uma ferramenta versátil, poderosa e extremamente útil na avaliação de desempenho. A sua principal vantagem é possibilitar que os modelos de simulação sejam construídos com níveis arbitrários de detalhes, permitindo simular situações complexas que são analiticamente intratáveis. Como desvantagens pode-se citar a complexidade no desenvolvimento do simulador e o tempo de execução da simulação. Um simulador pode utilizar números aleatórios para gerar variáveis aleatórias, que representam tempos de chegada e de serviços no sistema de acordo com distribuições de probabilidade. Com base nessas variáveis, questões de desempenho podem ser respondidas utilizando-se técnicas estatísticas para fornecer valores estimados. Para avaliar o desempenho de um sistema, uma vez construído o modelo probabilístico, asserções são feitas sobre o processo de chegada e o tempo de serviço de tarefas no sistema, as respostas às questões de desempenho podem, em teoria, ser analiticamente determinadas. Entretanto, na prática, estas questões são muito difíceis de serem determinadas analiticamente e as respostas a elas podem ser realizadas por um estudo de simulação [8]. Além disso, a representação por simulação do funcionamento de um determinado comportamento ou sistema é, na maioria das vezes, uma aproximação dos sistemas reais desenvolvida sinteticamente e que pode ser incompleta devido as simplificações realizadas na representação desse comportamento ou sistema. Assim, a confiabilidade dos resultados obtidos através dessa técnica pode ser, as vezes, variável ou até mesmo questionável.

3. Modelos analíticos - em sistemas computacionais muitas tarefas compartilham recursos tais como CPU, discos e outros dispositivos. Como, normalmente, somente uma tarefa por vez pode utilizar o recurso, todas as outras tarefas ficam esperando em filas por aquele recurso. O conjunto de recursos dá origem a uma rede de filas. Uma das formas mais conhecidas para a construção de modelos analíticos de filas é empregando a teoria de filas [7]. O sistema pode ser descrito por equações matemáticas, cujas soluções consistem na resolução do modelo. A teoria de filas é uma ferramenta matemática empregada para realizar análise de desempenho de um sistema o qual pode ser modelado como

uma rede de filas. Esta teoria ajuda a determinar, por exemplo, o tempo que as tarefas gastaram em várias filas dentro do sistema computacional. Estes tempos podem então ser combinados para predizer o tempo de resposta, o qual corresponde basicamente ao tempo total que a tarefa gastou dentro do sistema, incluindo o tempo de serviço. Além dessa ferramenta matemática pode-se citar outras, como: processos estocásticos, teoria da renovação, teoria dos grafos, geometria analítica, cálculo diferencial e integral, probabilidade e estatística, entre outras. Um modelo analítico pode ser aplicado em qualquer estágio do ciclo de vida de um sistema e o tempo exigido para sua aplicação é pequeno [7]. Essa ferramenta possibilita uma representação do funcionamento e/ou uma análise numérica do problema representado, permitindo assim a realização de uma avaliação de desempenho consistente e rápida. Além disso, quando o sistema a ser representado é muito complexo, essa técnica possibilita uma representação aproximada desse sistema, devido à necessidade de simplificações nessa representação para torná-la numericamente tratável. Geralmente, essas simplificações tornam a utilização da técnica analítica distante do comportamento real do sistema modelado e, portanto, pode ser necessária uma validação dos resultados, obtidos analiticamente, através das outras técnicas: simulação e/ou medição, descritos anteriormente.

Como foi já explicitado, uma modelagem analítica é muito importante para que se tenha uma adequada avaliação do protocolo da camada MAC do padrão IEEE 802.16. Dessa forma, o contexto dessa dissertação é a elaboração de um modelo analítico utilizando a teoria de filas e cadeia de Markov para representar o comportamento do protocolo da camada MAC do padrão IEEE 802.16 em termos do retardo fim-a-fim do sistema.

1.3 Objetivo do Trabalho

O objetivo do presente trabalho é avaliar o desempenho do protocolo MAC do padrão IEEE 802.16 em termos do retardo total. Para tal, será elaborado um modelo

analítico que represente as características do protocolo. Essa avaliação analítica é motivada pelo seguinte:

- A importância de se ter um método de avaliação de desempenho analítico para o protocolo MAC do padrão IEEE 802.16;
- A necessidade de verificar o nível de proximidade ou adequação dos modelos simulados utilizados nas pesquisas comparando-os com resultados obtidos analiticamente.

Para isso, será realizada uma análise, levando em consideração o comportamento dos dois extremos do protocolo de múltiplo acesso proposto pelo padrão IEEE 802.16. Para o esquema de alocação fixa, utiliza-se a teoria de filas e o esquema com acesso aleatório será modelado através de uma cadeia de Markov. Para a alocação dos dados e dos recursos pré-alocados para o canal compartilhado, a modelagem se dará através de um modelo *leaky-bucket* [1] com prioridade.

De forma mais objetiva, esse trabalho busca responder ou dar início a respostas para as seguintes perguntas em aberto:

1. Qual é o retardo fim-a-fim dos usuários de uma rede IEEE 802.16 onde existe a presença de diferentes tipos de tráfego?
2. Qual é o impacto dos diversos parâmetros do mecanismo de acesso aleatório no desempenho dos usuários que fazem uso dessa forma de acesso ao meio e da rede como um todo?
3. Qual o impacto do retardo do escalonamento das mensagens de dados, ou seja, após os recursos serem alocados no retardo total das mensagens da rede?
4. Qual o nível de proximidade ou semelhança do modelo analítico com os resultados de simulação?

Buscando-se alcançar esses objetivos e responder essas perguntas, na próxima seção as contribuições deste trabalho serão descritas.

[1] Sistema de enfileiramento com um único servidor que tem tempo de serviço constante.

1.4 Contribuições do Trabalho

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

- A elaboração de um modelo analítico, que leva em consideração as características do padrão IEEE 802.16 para usuários com diferentes requisitos de serviços para tipos de tráfego que requerem qualidade de serviço distintas;
- A investigação da influência de alguns parâmetros, tais como janela de contenção inicial e máxima e o número de retransmissões no mecanismo de acesso aleatório do protocolo MAC e no desempenho destas redes;
- A análise do retardo médio do escalonamento das mensagens com recursos pré-alocados;
- A avaliação de desempenho do padrão IEEE 802.16 sob a métrica do atraso total das mensagens;
- A validação da modelagem analítica com modelos simulados.

1.5 Organização do Trabalho

Para um melhor entendimento do restante deste trabalho, segue abaixo a estrutura da dissertação, indicando como esta encontra-se organizada.

No capítulo 2, é apresentado uma revisão dos conceitos básicos da teoria necessários ao entendimento desse trabalho. Além disso, são descritos os protocolos de múltiplo acesso e os aspectos da qualidade de serviço em redes sem fio.

No Capítulo 3, uma breve descrição do padrão 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. Finalmente, os trabalhos relacionados ao tema dessa dissertação são contextualizados em relação às redes metropolitanas sem fio.

O Capítulo 4 apresenta uma modelagem analítica para o retardo total das mensagens transmitidas no sub-canal de subida *uplink* para a avaliação de desempenho do padrão IEEE 802.16.

No Capítulo 5, 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 6 finaliza este trabalho consolidando os resultados apresentados no capítulo anterior através das conclusões e observações relevantes. Além disso, algumas perspectivas para trabalhos futuros são sugeridas.

Capítulo 2

Referencial Teórico

Este capítulo apresenta uma revisão dos conceitos básicos da teoria necessários ao entendimento desse trabalho. São descritos, em linhas gerais, os conceitos básicos dos protocolos de múltiplo acesso e os aspectos da qualidade de serviço em redes sem fio.

2.1 Protocolos de Múltiplo Acesso

Existe a necessidade de utilizar protocolos de múltiplo acesso para coordenar o compartilhamento do meio, quando um recurso é compartilhado por vários usuários independentes. Nesta dissertação, o recurso que deseja-se compartilhar é o canal de comunicação sem fio entre as estações. Nesse tipo de ambiente dinâmico, é necessário encontrar uma forma de compartilhar o canal de maneira adaptativa. Além das questões relacionadas à 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. Como exemplo, de custo, pode-se citar:

- colisões, devido a um fraco (ou nenhum) controle;
- capacidade de transmissão desperdiçada por causa de um controle muito rígido;
- sobrecarga adicional no tráfego do sistema em função de um controle dinâmico [10].

No intuito de permitir o acesso eficiente ao recurso disponível, existe desperdício devido ao custo de organizar as demandas em algum tipo de fila cooperativa.

Devido ao exposto acima, 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 da 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. 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.

Muitas questões envolvendo redes de computadores tratam da alocação eficiente do canal de comunicação entre demandas competitivas. Existe uma grande vantagem no tratamento ao compartilhamento de recursos. Por exemplo, há duas soluções para

o acesso ao canal de comunicação: a solução clássica provê um canal dedicado para cada usuário que necessita acessar o meio pelo tempo que for necessário; a outra solução 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.

Classificação

Devido a grande variedade de funcionalidades em relação a alocação estática ou dinâmica do canal, do mecanismo de controle centralizado ou distribuído e o comportamento adaptativo do algoritmo de controle existentes nos diversos protocolos de controle de acesso ao meio (*Media Access Control* - MAC) estes protocolos podem ser classificados de acordo com a figura 2.1 [11] [1]. Não há nenhum protocolo que se sobressaia aos outros sob todos os aspectos de desempenho, cada classe tem suas próprias vantagens e desvantagens.

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

- Protocolos de Alocação Fixa (canal ocioso) - 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 mensagens para enviar durante o período de tempo em que o meio está alocado para ela, o canal ficará ocioso, ou seja, sem transmissão de dados enquanto outros terminais poderiam utilizá-lo.

- Protocolos de Acesso Aleatório (colisões) - Os protocolos de acesso aleatório não possuem um controle rígido para alocação do canal, também são relativamente simples de implementar, porém, a possibilidade de ocorrer colisões

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

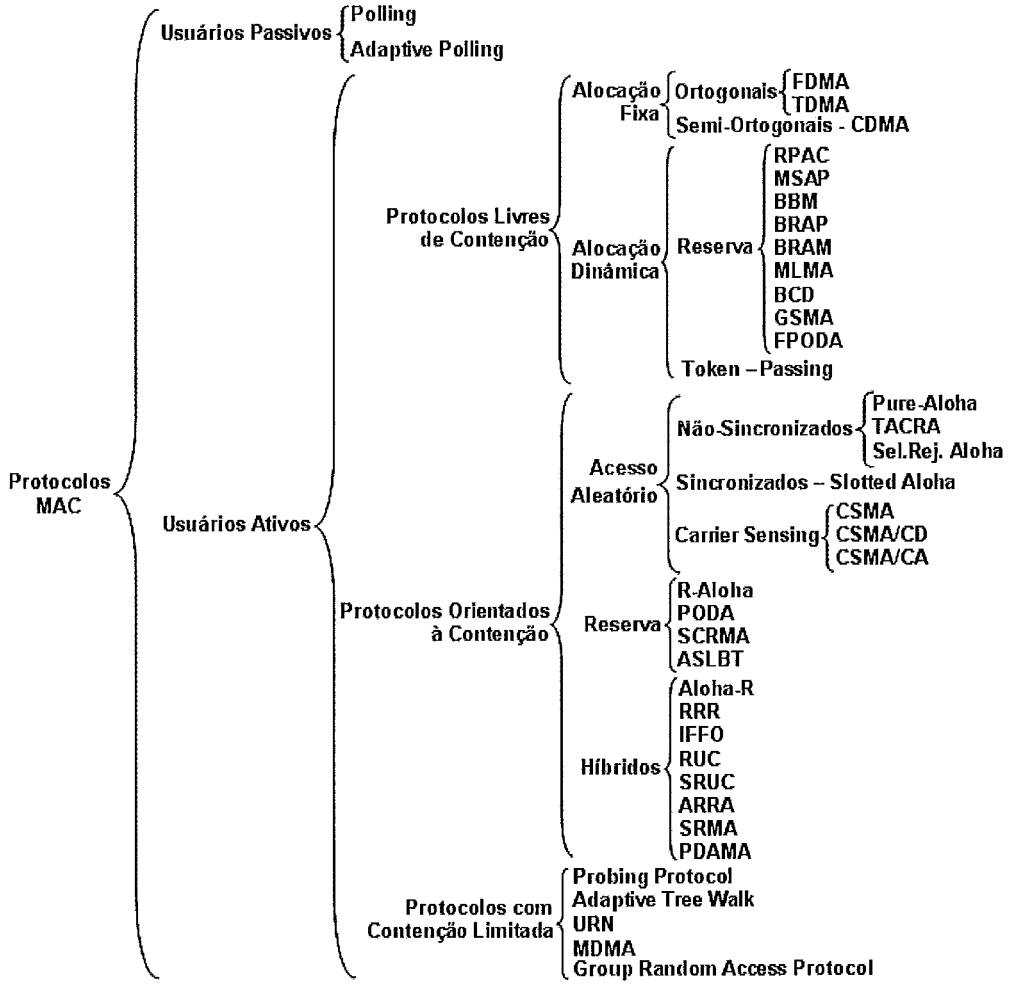


Figura 2.1: Classificação dos protocolos MAC.

quando duas ou mais estações tentam transmitir ao mesmo tempo provoca desperdício no canal de transmissão. Pode-se citar o ALOHA [16] e o CSMA (*Carrier Sense Multiple Access*) [17] como exemplos desta categoria.

- Protocolos de Alocação Dinâmica (sobrecarga) - Nos protocolos de alocação dinâmica o canal é alocado de acordo com as necessidades das estações. Existem algumas formas de implementar este controle, 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 há uma sobrecarga na rede devido aos sinais de controle. Porém, não existe a possibilidade de colisões e o canal é alocado sob demanda, evitando períodos de

tempo ociosos no canal.

Contudo, 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.

Uma característica interessante é que alguns protocolos de múltiplo acesso já incorporam mecanismos de escalonamento das mensagens através de prioridades, possibilitando uma diferenciação no tratamento de classes de tráfego distintas. Esta diferenciação é fundamental para transmissão de dados e serviços multimídia com diferentes requisitos de qualidade de serviço (QoS). Dentro desse contexto será investigada, nessa dissertação, esta característica no protocolo MAC do padrão IEEE 802.16 que será apresentado no capítulo 3.

2.2 Qualidade de Serviço

PARA dar suporte à grande diversidade de aplicações disponíveis na Internet, tais como serviços de voz, vídeo, multimídia e transferência de arquivos, garantir a qualidade de serviço (QoS) em redes de computadores tornou-se uma necessidade básica. Para que essa tarefa seja realizada eficientemente, deve-se projetar e implementar um conjunto de mecanismos que incluem policiamento, moldagem do tráfego, controle de admissão e escalonamento. Essa seção apresenta os conceitos básicos inerentes à qualidade de serviço em redes de computadores.

Algumas aplicações são relativamente insensíveis à degradação transitória da qualidade de serviço oferecida pela rede. Por exemplo, num serviço de transferência de arquivos, a redução da largura de banda disponível ou o aumento do atraso dos pacotes podem afetar o desempenho da aplicação, mas não comprometem a sua operação. Devido à capacidade de adaptação e as variações na disponibilidade de recursos, essas aplicações contentam-se com um serviço do tipo melhor esforço, no qual a rede compromete-se apenas em tentar transmitir o tráfego gerado pela aplicação, sem no entanto oferecer garantias de desempenho. Já no caso de aplicações

em tempo real, a diminuição da largura de banda disponível ou o aumento do atraso podem inviabilizar a sua operação. Neste caso, a rede necessita reservar recursos para as aplicações de modo que, mesmo em momentos de maior carga na rede, os requisitos mínimos de desempenho destas aplicações sejam atendidos, ou seja, a rede deve fornecer um serviço com garantias de qualidade de serviço (QoS).

A forma de se obter qualidade de serviço, envolve a inclusão de mecanismos que buscam racionalizar o uso dos recursos disponíveis na rede. Esses mecanismos estabelecem níveis de serviço e permitem a convivência na mesma rede de tráfegos com requisitos distintos de qualidade. Tráfegos pertencentes a níveis de serviço diferentes são tratados de forma que o nível mais prioritário possa sempre dispor dos recursos de que necessita, ainda que em detrimento dos níveis menos prioritários. Ao mesmo tempo, tráfegos pertencentes a um mesmo nível de serviço são tratados de maneira que suas demandas sejam atendidas de forma justa.

Dentre as métricas de QoS mais utilizadas na literatura destacam-se:

- retardo médio - 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;
- *jitter* - variação do atraso. Aplicações com taxa de bit constante são bastante suscetíveis ao *jitter*;
- taxa de perda - 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;
- banda obtida - Da mesma forma, algumas aplicações necessitam de uma quantidade mínima de largura de banda para transmitir dados.

Em [20], os autores identificam quatro princípios básicos para prover qualidade

de serviço em aplicações multimídia:

- Classificação de pacotes - 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 distingüi-los;

- Escalonamento e policiamento de tráfego - trata a diferenciação no tratamento dos pacotes de diferentes classes de tráfego. 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 sejam 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;

- Eficiência - A eficiência de um protocolo de controle de acesso ao meio é uma medida do aproveitamento da largura de banda disponível, sendo normalmente expressa pela razão entre a taxa útil e a capacidade do canal. Um protocolo de controle de acesso ao meio deve procurar maximizar a eficiência sem comprometer a qualidade de serviço oferecida às conexões. Para isso, deve procurar minimizar o *overhead* que introduz.

- Controle de admissão - restringe o número de usuários simultaneamente presentes na rede de forma a evitar a saturação do enlace sem fio e, conseqüentemente, a violação dos requisitos de QoS.

As seções seguintes provêem uma visão geral de vários mecanismos de implementação dos quatro princípios listados acima, considerando as principais disciplinas de filas utilizadas.

2.2.1 Tipos de Tráfego

O primeiro passo para prover diferentes níveis de serviços, para suportar as aplicações multimídia, é através da classificação de pacotes. 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 (*Cosntant Bit Rate* - CBR). Entretanto, esquemas de compressão convertem este tipo de tráfego para uma taxa variável de bits (*Variable Bit Rate* - VBR). Esse tipo de tráfego 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.

2.2.2 Escalonamento de Pacotes

A disciplina de escalonamento diz respeito à política de transmissão de pacotes utilizada. Uma das principais formas de prover qualidade de serviços numa rede de computadores é incorporar disciplinas de escalonamento ao protocolo de acesso ao meio, para que seja obtida a diferenciação de serviços na camada MAC. Dentro desse contexto, serão apresentadas duas importantes disciplinas de escalonamento de pacotes estudadas na literatura: *first-in-first-out* e fila de prioridades.

First-In-First-Out (FIFO)

A disciplina ***First-In, First-Out*** **(FIFO)** é a abordagem mais simples para gerenciar o escalonamento de pacotes, onde todos os pacotes que chegam são colocados em uma fila comum e servidos pela ordem de chegada, como ilustra a Figura 2.2. Quando a fila está cheia, ocorre o descarte de pacotes (*packet loss*). Além disso, não é possível prover diferentes níveis de QoS para fluxos distintos, visto que a disciplina FIFO trata todos os pacotes de maneira igual. Resta ainda comentar que quando um usuário envia pacotes, utilizando essa disciplina, a uma alta taxa, ele ocupa todo o sistema, impedindo outros usuários de acessá-lo - *hogging* [22].

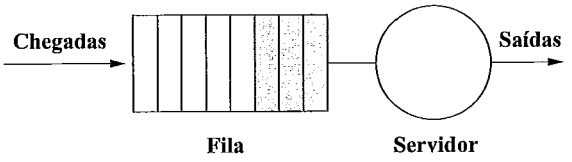


Figura 2.2: Abstração de um fila FIFO.

Fila de Prioridades (HOL - *Head-Of-the-Line*)

Nesta disciplina, uma fila separada é mantida para cada classe e os pacotes que chegam ao sistema são classificados em duas ou mais classes de prioridade, sendo transmitido sempre o pacote que estiver na "cabeça" da fila de maior prioridade, que não se encontra vazia. E os pacotes que pertencem a mesma classe de prioridade, podem ser servidos como uma disciplina FIFO. A figura 2.3, ilustra a fila de prioridades.

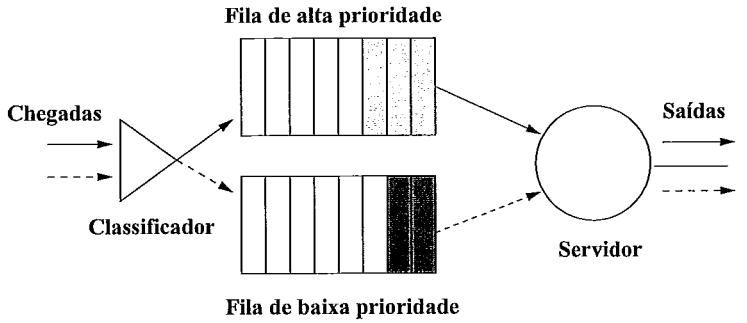


Figura 2.3: Modelo de fila com prioridades.

2.2.3 Controle de Admissão

O controle de admissão tem como principal função decidir adequadamente se um canal de comunicação pode ou não aceitar uma nova conexão. A nova conexão será aceita, se a qualidade de serviço para todas as fontes que compartilham o mesmo canal (incluindo a nova) for satisfeita, 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.

Cada fonte deve especificar o seu fluxo através de um conjunto de parâmetros conhecidos como **descritores de tráfego** para que este controle determine se a QoS no sistema será mantida e para poder calcular a quantidade de banda que deve ser reservada para cada recurso. Estes descritores devem caracterizar os fluxos de tráfego de maneira compacta e eficiente.

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. Existe uma grande variedade de mecanismos propostos na literatura para controle de admissão e para um estudo mais detalhado sobre o assunto recomenda-se [23].

2.2.4 Policiamento de Tráfego

Como foi estudado, uma 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 balde furado (*leaky bucket*) 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. O balde furado consiste em uma fila finita. Quando chega um pacote, se houver espaço na fila, ele é incluído na fila, caso contrário, ele é descartado. A cada unidade de tempo, um pacote é transmitido (a menos que a fila esteja vazia). Pode-se ainda, utilizar este mecanismo em conjunto com a disciplina de escalonamento HOL vista anteriormente [20].

2.3 Considerações Finais

Neste capítulo foram introduzidos os principais conceitos relacionados a essa dissertação. Onde foram apresentadas as principais características dos protocolos de múltiplo acesso e os principais conceitos referentes a qualidade de serviço em redes sem fio foram abordados. No próximo capítulo será apresentado o padrão IEEE 802.16 e os trabalhos relacionados ao tema dessa dissertação.

Capítulo 3

Padrão IEEE 802.16 e Trabalhos Relacionados

ESTE capítulo descreve de maneira abrangente os aspectos significativos do padrão IEEE 802.16 e realiza 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. Além disso, serão apresentados os trabalhos relacionados ao tema dessa dissertação.

3.1 Visão Geral

O padrão IEEE 802.16 [4] especifica uma interface sem fio para redes metropolitanas (WMAN) e vem sendo desenvolvido com a finalidade de padronizar a tecnologia de acesso sem fio à banda larga. O padrão define a interface aérea e o protocolo de acesso ao meio para redes metropolitanas sem fio fornecendo altas taxas de transmissão para o acesso comercial e residencial à Internet.

Para promover e certificar a compatibilidade e interoperabilidade entre os equipamentos de acesso sem fio à banda larga, que estejam em conformidade com o padrão IEEE 802.16, foi formado o WiMAX Forum. O WiMAX (*Worldwide Interoperability for Microwave Access*) foi definido pelo WiMAX Forum como uma tecnologia que possibilite uma alternativa a conexão a cabo ou DSL (*Digital Subscriber Line*). Esse fórum tem como função certificar que os equipamentos industriais e produtos comerciais estejam em conformidade entre si, além de promover o uso desta tecnologia.

A arquitetura de uma rede que utiliza o padrão IEEE 802.16 possui dois elementos principais: Estação Base (*Base Station* - BS) e Estação Cliente (*Subscriber Station* - SS), como mostra a Figura 3.1. A BS é o nó central que coordena toda a comunicação e as SSs se localizam a diferentes distâncias da BS, em uma topologia Ponto-Multiponto (PMP). 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.

O padrão também permite topologia *Mesh* (opcional). A principal diferença entre essas topologias, *Mesh* e PMP, está no fato de que em uma rede PMP o tráfego flui apenas entre a BS e as SSs, enquanto que no modo *Mesh*, o tráfego pode ser roteado através das SSs e pode ocorrer diretamente entre duas SSs, como mostra a figura 3.2. Este trabalho concentra-se nas redes com topologia PMP.

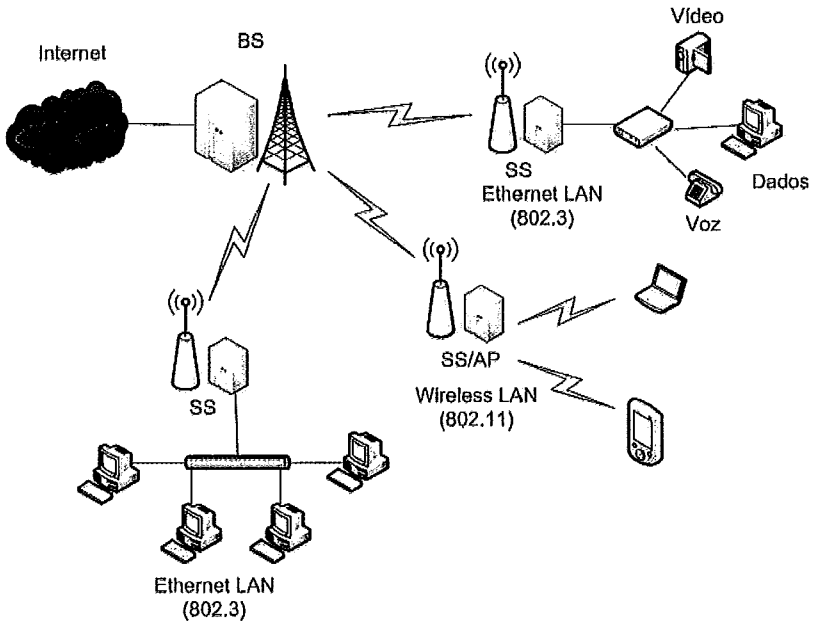
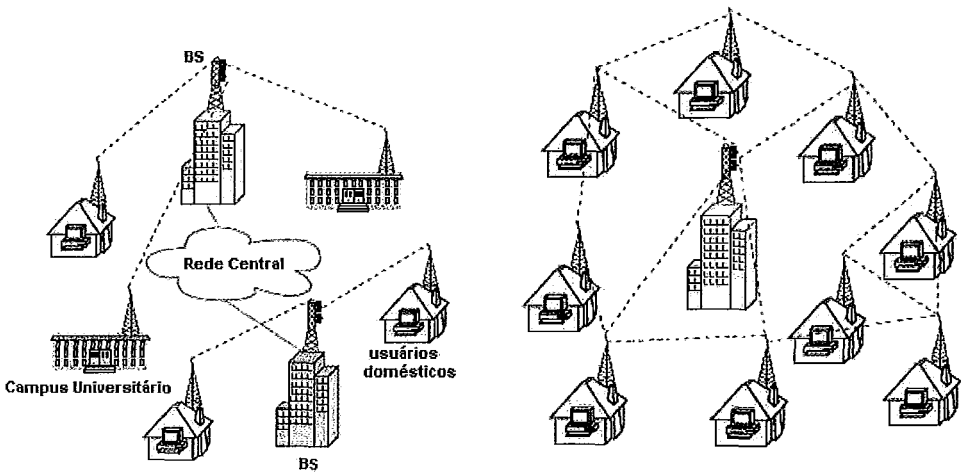


Figura 3.1: Arquitetura básica do sistema BWA.

Esta tecnologia foi desenvolvida para alavancar o acesso sem fio à banda larga em redes metropolitanas (MANs), oferecendo desempenho comparável às tradicionais tecnologias a cabo e DSL. Entretanto, as principais vantagens desta tecnologia 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.

3.2 Camada PHY e MAC

Com o intuito de especificar formalmente as redes sem fio banda larga que cobrissem áreas metropolitanas, em 1999, foi criado pelo *Institute of Eletrical and Eletronics Engineers* o *Broadband Wireless Access* (BWA) *Working Group*, ou IEEE 802.16. Em sua primeira versão, em 2001, o padrão IEEE 802.16 operava em um intervalo de freqüência licenciada entre 10 e 66 GHz, onde era necessário o uso de antenas direcionais para obter desempenho satisfatório. Na área metropolitana, entretanto, não se pode assegurar a operação com linha de visada, devido a presença



(a) Ponto-Multiponto (b) Mesh

Figura 3.2: Topologias permitidas pelo padrão IEEE 802.16

de obstáculos, como: edificações, árvores, etc. Assim, o padrão foi estendido com as versões 802.16a, 802.16b e 802.16c, abordando, respectivamente, problemas relacionados ao espectro de freqüências, a qualidade de serviço e a interoperabilidade do padrão. Outras emendas foram autorizadas posteriormente, o 802.16e fornece suporte a mobilidade, o 802.16f tem o objetivo de melhorar a funcionalidade de múltiplos saltos e o 802.16g prevê melhorar o *handover* e a QoS.

A tecnologia WiMAX pode alcançar, teoricamente uma área de cobertura de 50 Km [24]. As taxas de transmissão de dados vão de 50 à 150 Mbps, dependendo da largura de freqüência do canal e do tipo de modulação [25]. 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 *downlink*, os dados são transmitidos por difusão, enquanto no *uplink* o meio é compartilhado através de múltiplo acesso. A duplexação entre eles pode ser feita de duas maneiras: duplexação por divisão de freqüências (*Frequency-Division Duplexing* - FDD) e duplexação por divisão do tempo (*Time-Division Duplexing* - TDD). Basicamente, no FDD, os dois canais compartilham o mesmo tempo e os dados são transmitidos em freqüências diferentes, como a transmissão no canal de *downlink* pode ser feita em rajadas, existe suporte

a estações tanto *full-duplex* quanto *half-duplex*. Já no TDD, os canais são divididos no tempo, e utilizam a mesma freqüência. Embora o quadro possua tamanho fixo na duplexação por tempo, a divisão entre os tempos fornecidos para *downlink* e para o *uplink* pode ser desigual. As Figuras 3.3 e 3.4 mostram a estrutura do quadro PHY com TDD e FDD.

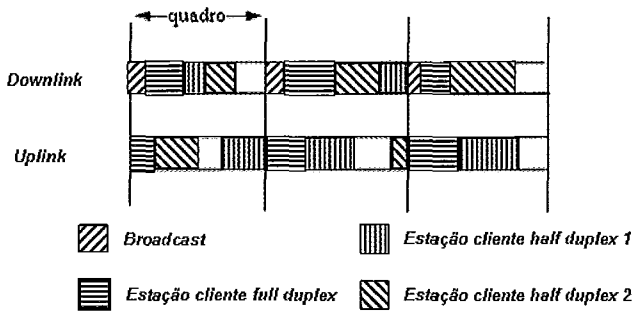


Figura 3.3: Estrutura do Quadro FDD

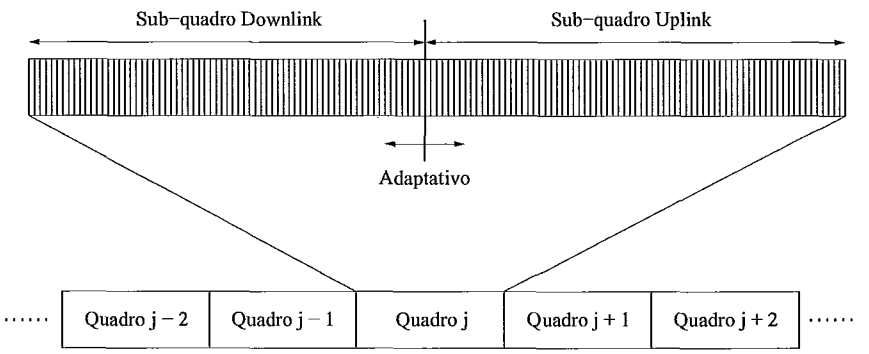


Figura 3.4: Estrutura do Quadro TDD

Como foi visto, na duplexação por divisão de freqüência, FDD, o *downlink* suporta estações cliente *full-duplex* e *half-duplex*, e é possível que estas últimas estejam programadas para transmitir posteriormente no mesmo quadro. Há uma seção de controle de quadro, que possui os DL-MAP e UL-MAP, utilizados para indicar os segmentos físicos nos quais as rajadas devem começar, e uma seção TDMA (*Time-Division Multiple Access*), controlada pelo DL-MAP, que permite a decodificação de somente algumas regiões específicas do sub-quadro, o que é utilizado por estações clientes *half-duplex* que necessitem transmitir antes de receber o *downlink* completo.

Na duplexação por divisão no tempo, durante o *downlink* os pacotes de dados são transmitidos por difusão pela BS para todas as SSs, que por sua vez, capturam apenas os pacotes destinados a elas, portanto, a transmissão é relativamente simples pois somente a BS transmite neste sub-quadro. Durante o *uplink*, através da mensagem UL-MAP no começo de cada quadro, a BS transmite por difusão o número de segmentos que será atribuído para cada SS dentro do sub-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 *uplink*. Portanto, 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 *uplink* para determinar as oportunidades de transmissão (IEs) utilizando as requisições (BW-Request) enviadas pelas SSs. A Figura 3.5 ilustra a estrutura do quadro MAC no esquema de alocação TDD.

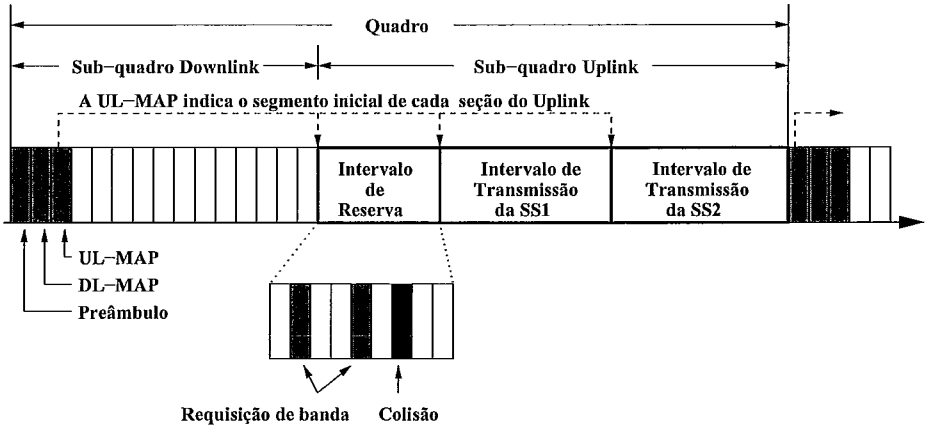


Figura 3.5: Estrutura do quadro MAC no esquema TDD.

Para enviar requisições de oportunidades de transmissão para a BS, as SSs [4] utilizam acesso aleatório e *piggybacking*[1] no sub-quadro de *uplink*. A BS é responsável por estabelecer um intervalo de reserva no início do sub-quadro de *uplink* para que as SSs possam requisitar as oportunidades de transmissões no próximo sub-quadro de *uplink*, ou em algum mais a frente, dependendo da ocorrência ou não de colisões. É importante notar que o IEEE 802.16 utiliza um protocolo de acesso ao meio ba-

[1]Requisições enviadas pelas SSs no final do quadro de dados, transmitidas durante o *uplink*.

seado 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 *uplink* mais a frente, como indica a Figura 3.6. Para a resolução de colisões neste intervalo, o padrão define o algoritmo *binary truncated exponential backoff*, onde uma SS detecta a ocorrência de colisão caso a UL-MAP do próximo quadro não contenha nenhuma oportunidade de transmissão destinada a ela.

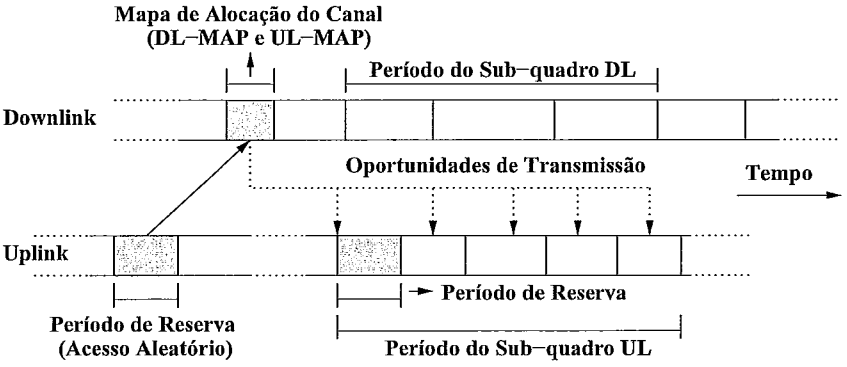


Figura 3.6: Estrutura de alocação do IEEE 802.16.

A camada de acesso ao meio (*Medium Access Control layer - MAC*) é orientada à conexão. Cada conexão é identificada por um identificador (*Connection Identifier - CID*) de 16 bits e cada SS tem um endereço MAC único que a identifica e é utilizado para registrá-la e autenticá-la na rede. Todo tráfego, incluindo o tráfego não orientado à conexão, é mapeado para uma conexão. Além do gerenciamento das conexões, a camada MAC é responsável pelo controle de acesso ao meio e pela alocação de banda.

A alocação dos recursos para as SSs é realizada sob demanda. Quando uma SS precisa de largura de banda para uma conexão, ela envia uma mensagem de requisição para a BS. Uma requisição de banda pode ser enviada como um pacote individual em um *grant* (garantia da oportunidade de transmissão) reservado para esse fim, ou pode ser enviada juntamente com um pacote de dados (*Piggybacking*). A requisição de largura de banda pode ser incremental, indicando a largura de banda adicional que a estação precisa, ou agregada, indicando a largura de banda total requisitada pela SS. Para a SS, as requisições de banda sempre são referentes a uma

determinada conexão, enquanto os *grants* alocados pela BS são destinados à uma SS e não a uma conexão particular. Dessa forma, a SS pode utilizar o *grant* recebido para uma conexão diferente daquela para a qual a requisição foi feita. A alocação de *grants* para o envio de requisições de banda, chamada *polling*, pode se dar de duas formas: *unicast*, onde a SS recebe um *grant* cujo tamanho é suficiente para o envio de uma requisição de banda; e baseado em contenção, onde a BS aloca um *grant* para um grupo de SSs, as quais devem competir pela oportunidade de enviar a mensagem de requisição. Para reduzir a probabilidade de colisão, apenas as SSs que necessitam de banda participam da contenção. Para resolução da contenção, as estações devem utilizar o algoritmo de *backoff* exponencial. O tamanho da janela mínima e da janela máxima de contenção é controlado pela BS.

3.3 Qualidade e Escalonamento de Serviços

A camada MAC também provê mecanismos para fornecer QoS aos diferentes tipos de tráfego. O principal mecanismo para provisão de QoS consiste em associar os pacotes que passam pela camada MAC a um fluxo de serviço. O fluxo de serviço é um serviço da camada MAC que fornece transporte uni-direcional as mensagens. Durante a fase de estabelecimento da conexão, esses fluxos de serviço são criados e ativados pela BS e pela SS. Cada fluxo de serviço deve definir seu conjunto de parâmetros de QoS, dentre eles retardo máximo, largura de banda mínima e o tipo de serviço de escalonamento.

O padrão especifica quatro serviços de escalonamento, onde cada fluxo é associado a um desses serviços e o escalonador da BS aloca largura de banda para as SSs seguindo o conjunto de regras definido por eles [4]:

- *Unsolicited Grant Service* (UGS): suporta fluxos de tempo real que geram pacotes de dados, com tamanho fixo, periodicamente, tal como voz sobre IP . O serviço oferece *grants* com tamanho fixo periodicamente. Fluxos UGS não podem utilizar segmentos de contenção reservados para requisição de largura

de banda.

- *Real-Time Polling Service* (rtPS): suporta fluxos de tempo real que geram pacotes, com tamanho variável, periodicamente, como por exemplo vídeo MPEG. O serviço oferece segmentos *unicast* periódicos para requisição de largura de banda, os quais satisfazem as necessidades do fluxo e permitem à SS especificar o tamanho desejado para o *grant*. Conexões rtPS não podem utilizar segmentos de contenção reservados para requisição de largura de banda.

- *Non-Real-Time Polling Service* (nrtPS): Suporta tráfego que são insensíveis ao atraso no tempo e requerem um mínimo de alocação de banda. 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 arquivos. Conexões nrtPS podem utilizar segmentos de contenção reservados para requisição de largura de banda. Os *grants* alocados para esse serviço ocorrem com menor freqüência que para o rtPS.

- *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. Conexões BE utilizam segmentos de contenção reservados para requisição de largura de banda.

Como foi visto, os diferentes serviços de escalonamento, especificados no padrão, são bem definidos para diferentes aplicações e com respeito a diferentes tipos de tráfego e requisitos de QoS. Um problema que diz respeito a melhoria dos serviços de escalonamento, UGS e rtPS, é que para usá-los eficientemente, é necessário um bom conhecimento das características do tráfego. Por exemplo, para transportar uma conexão VoIP, via UGS, é necessário conhecer o codec de voz, isto é, o tempo entre-chegadas dos pacotes e o tamanho do pacote, para garantir o recurso para as estações eficientemente. Visto que as características do retardo de muitas aplicações como jogos, *uploads* de *webcam*, chamadas *skype*, que se ajustariam bem a uma das melhorias dos serviços de escalonamento, não são conhecidas ou não sinalizadas automaticamente para a BS, assim, essas aplicações são tipicamente transporta-

das via nrtPS ou BE. Isto faz o acesso aleatório simples um dos mais importantes mecanismos para o desempenho do WiMAX.

Ainda resta comentar que, como dito anteriormente, do ponto de vista do mecanismo MAC, a diferença entre os serviços nrtPS e BE é que a BS fornece *polls* unicast periódicos para o serviço nrtPS numa escala de tempo de 1 seg. Assim, para uma alta carga, as estações requisitam largura de banda via *piggybacking*, por outro lado, para o sistema com carga baixa, a escala de tempo para poll unicast é maior do que o tempo necessário para requisitar largura de banda usando contenção. Logo, a diferença entre o serviço BE e nrtPS é desprezível.

3.4 Trabalhos Relacionados

Através da seção anterior identifica-se que o padrão IEEE 802.16 utiliza um protocolo de acesso ao meio baseado em alocação dinâmica com reserva. Neste capítulo, alguns trabalhos relacionados ao desempenho do padrão IEEE 802.16 propostos na literatura, são contextualizados em relação a redes metropolitanas sem fio.

A análise de desempenho do padrão IEEE 802.16 foi extensivamente estudada na literatura [26, 27, 28, 29, 30, 31, 32]. Porém, esses trabalhos desenvolvem avaliação de desempenho através de resultados obtidos por simulação. Além disso, nenhum modelo analítico para o retardo total das mensagens foi apresentado, buscando representar de forma exata alguma métrica de desempenho do protocolo de acesso ao meio (como atraso ou vazão).

Em [26], os autores apresentam um estudo simulado do protocolo MAC do IEEE 802.16 operando com interface aérea OFDM e com estações *full-duplex*. Eles avaliam o desempenho do sistema em diferentes cenários de tráfego e variando os valores de um conjunto relevante de parâmetros do sistema. Com respeito ao tráfego de dados, concluíram que há um *trade-off* entre o retardo médio e a vazão em relação à duração do quadro. Além disso, foi observado que a sobrecarga devido à transmissão dos preâmbulos físicos aumenta com o número de estações. Finalmente, mostra-se que as estações estão aptas a requisitar largura de banda no *uplink* para BS eficientemente

usando requisições de largura de banda por *piggybacking*. Além disso, eles mostram que o desempenho das conexões, em termos do retardo, é fortemente dependente do retardo introduzido pelo mecanismo de requisição de largura de banda.

Em [29], Aura Ganz *et al.* propõem um escalonador de pacotes para o sub-canal *uplink* 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 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 [32], Dong-Hoon Cho *et al.* propõem uma nova arquitetura de QoS para o padrão IEEE 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, o trabalho não especifica claramente como é calculado o tempo de vida de cada pacote.

Em [30], 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 [32], onde o melhor tamanho para a janela de contenção é igual ao número de estações ativas.

Em [27], é estudado o desempenho de diferentes mecanismos de acesso aleatório presente no padrão IEEE 802.16, via simulação. Mostra-se que o montante das fontes que serão reservadas para o acesso aleatório depende fortemente do desempenho desejado. Se um maior retardo de acesso pode ser tolerado, uma maior vazão pode ser alcançada. Se entretanto, menores retardos são requeridos por alguns usuários que transportam tráfego sensível ao retardo, através de conexões *best-effort*, uma parte considerável do subquadro de *uplink* será desperdiçada para o acesso aleatório.

Em [28], é investigado o desempenho do subquadro de *upstream* do protocolo MAC do padrão IEEE 802.16, via resultados simulados, em termos da vazão, retardo médio e probabilidade de colisão com respeito ao intervalo MAP, a taxa entre segmentos de contenção e segmentos de dados, para estimar o uso eficiente da largura de banda no subquadro de *uplink*. É investigado como o desempenho varia como uma função da carga da rede para diferentes tamanhos de *payload* e número de estações. Além disso, através dos resultados obtidos, foi observado que a utilização da largura de banda do *upstream* é melhor quanto maior é o tamanho do *payload* devido a redução do overhead comparado ao overhead associado aos pacotes de tamanho menores. O retardo médio aumenta para um grande número de estações, devido ao aumento das colisões. Também é observado que, a probabilidade de colisão varia como função da carga da rede para um dado número de estações.

Em [31], os autores propõem um protocolo MAC baseado em *polling* e apresenta um modelo analítico para avaliar o seu desempenho em termos do retardo de pacotes, considerando um sistema onde a BS interroga cada nó em cada quadro para determinar seus requerimentos de largura de banda, com três estratégias de escalonamento diferentes. Foram desenvolvidas expressões analíticas de forma fechada para o retardo nos casos onde as estações são interrogadas no início ou no fim do subquadro de *uplink*.

Novamente, é importante notar que nenhuma das propostas apresentadas nesta seção envolvem a análise de desempenho para o retardo total das mensagens do protocolo de acesso ao meio do padrão IEEE 802.16. Além disso, alguns trabalhos fazem a análise com a adição de algum componente de hardware e/ou camada de software para especificar uma solução para o escalonamento de pacotes.

3.5 Considerações Finais

Neste capítulo foi descrito o padrão IEEE 802.16, bem como sua camada física e MAC, além de seus mecanismos de QoS. Além disso, foram apresentados os trabalhos relacionados ao tema dessa dissertação. No próximo capítulo, será visto como a solução proposta neste trabalho busca introduzir um modelo analítico que busca incorporar as características do protocolo MAC do IEEE 802.16. Dessa forma, pode-se avaliar o desempenho dessas redes, sob a métrica de desempenho do retardo total das mensagens e do retardo das mensagens de requisição de largura de banda, proporcionando assim, um primeiro passo para a especificação de uma modelagem completa dos mecanismos de escalonamento para as redes metropolitanas sem fio.

Capítulo 4

Modelo Analítico Proposto

PARA avaliar o desempenho do padrão IEEE 802.16, este capítulo apresenta uma modelagem analítica para o retardo total das mensagens transmitidas no sub-canal de *uplink*. Considera-se dois modelos separados: um para a fase de requisição de largura de banda (resolução de colisão) e outro para a transmissão de dados. Primeiramente, será detalhada a análise para a fase de requisição de largura de banda e em seguida a análise para a transmissão de dados é apresentada. Por fim, são realizadas as análises para o retardo total das mensagens.

4.1 Visão Geral

Em qualquer sistema, computacional ou não, seu bom desempenho é um fator que deve ser buscado durante todo o seu desenvolvimento. Sendo assim, é necessária uma forma de medir ou dimensionar o desempenho desses sistemas. Para realizar esse dimensionamento são utilizadas ferramentas para medição e análise de desempenho, que disponibilizam ao analista medidas e métricas distintas. Nesse escopo, serão apresentadas algumas características discutidas anteriormente que serão discutidas e relacionadas ao proposto nesta dissertação.

Como foi visto no capítulo anterior, as redes IEEE 802.16 são reguladas por um mecanismo de requisição-garantia no processo de requisição de largura de banda, e por garantias de recursos pré-negociados para o tráfego mais prioritário. Sendo assim, a modelagem se fará em etapas, bem definidas, onde serão incorporadas as seguintes características ao modelo:

Cadeia de Markov bi-dimensional - é nosso ponto de partida. A cadeia de Markov representa o processo para a resolução de contenção, das mensagens de requisição de largura de banda, do processo de requisição-garantia.

Fila de requisição e fila de dados - Um usuário primeiro envia uma mensagem de requisição para a BS e uma vez que essa requisição é garantida, é permitido ao usuário enviar a mensagem real. Ou seja, no momento em que o usuário gera uma mensagem de requisição, ele entra na fila de requisição, quando a requisição é garantida, o usuário migra da fila de requisição para a fila de dados, onde ele espera até ter a oportunidade de transmitir a mensagem de dados, conforme mostra a figura 4.1.

Escalonamento centralizado - O escalonador da BS determina de que forma a capacidade do canal de *upstream* é dividida entre as SS. Usando a abstração das filas de requisição e de dados, a capacidade do *upstream* é dividida nessas duas filas.

Requisição virtual - Um usuário que tem tráfego de tempo real, envia uma mensagem de requisição virtual, a qual terá prioridade sobre as demais, para esca-

lonar recursos na rede.

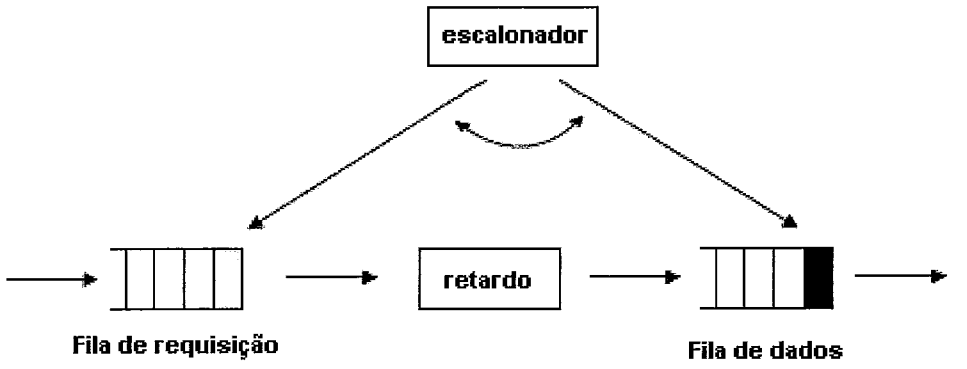


Figura 4.1: Escalonamento das mensagens reguladas pelo mecanismo de requisição/garantia

Nas próximas seções será realizada a modelagem analítica proposta neste capítulo, visando abranger a análise do retardo total das mensagens de dados, considerando o retardo de reserva, ou seja, na fase de requisição de largura de banda e o retardo de transmissão, ou seja, na fase de alocação das mensagens de dados.

4.2 Retardo na Fase de Requisição de Largura de Banda

Como foi estudado, colisões podem ocorrer quando uma SS envia uma mensagem de requisição de largura de banda para a BS no período de contenção. No IEEE 802.16 o algoritmo de *backoff* é utilizado para minimizar a ocorrência de tais colisões [4]. A técnica utilizada para obter o retardo médio das mensagens de requisição de largura de banda é similar ao método empregado em [33].

4.2.1 Visão Geral do Princípio Operacional do Algoritmo de *Backoff*

O método de resolução de contenção do padrão IEEE 802.16 é baseado no *backoff* exponencial binário truncado, com a janela de contenção inicial e máxima controladas pela BS. Os valores são especificados no UL-MAP e representam um valor com potência de 2. Por exemplo, um valor de 4 indica uma janela de contenção entre 0 e 15.

Quando uma estação tem novo dado a transmitir, ela ajusta o parâmetro expoente de *backoff* (BE) igual ao valor que fornece a janela de contenção mínima. A estação seleciona um número aleatório entre 0 e $2^{BE} - 1$. Este valor aleatório indica o número de oportunidades de transmissão de contenção que a SS deverá aguardar para transmitir a mensagem de requisição de largura de banda. Depois de transmitir a mensagem de requisição de largura de banda, no período de contenção, a estação espera por um *grants* no MAP subseqüente, uma vez recebido, a resolução do processo de contenção está completa. Se a mensagem de requisição de largura de banda colide, a SS executa os seguintes passos: o valor do número de tentativas de *backoff* (N_B) é acrescido de 1 unidade. Se BE é menor que m (estágio máximo de *backoff*), então BE é também acrescido de 1 unidade. Finalmente ocorre a seleção aleatória anterior, e os passos de adiamento são repetidos. O algoritmo é repetido até a mensagem de requisição de largura de banda ser transmitida com sucesso ou até o número máximo de tentativas ser alcançado $(m+R)$ e o pacote ser descartado [4].

4.2.2 Modelagem e Análise do Algoritmo de *Backoff*

Nesta seção será construído um modelo do algoritmo de *backoff* e discutido o comportamento de uma única estação com o modelo de Markov. Será obtida a probabilidade estacionária, τ, de que a SS transmita uma mensagem num segmento de tempo genérico, isto é, escolhido aleatoriamente. Este parâmetro é usado para fornecer o retardo médio das mensagens de requisição de largura de banda.

Considere um número fixo de N estações em contenção. Em condições de saturação, cada estação tem imediatamente um pacote disponível para transmissão, após cada transmissão completada com sucesso. Esse hipótese é necessaria para garantir que todas as estações participam da disputa pelo acesso ao meio, e permitir a subsequente derivação de τ.

Seja $B(t)$ o processo estocástico representando o tamanho do contador do tempo de *backoff* para uma determinada estação. O contador de *backoff*, k é definido como o número de oportunidades de transmissão para o qual uma estação deve esperar antes de iniciar a sua transmissão. Este processo estocástico é não Markoviano, visto que possíveis valores do contador de *backoff* de cada estação dependem da sua história de transmissão particular (por exemplo, quantas retransmissões o pacote na cabeça da fila sofreu). Uma escala de tempo discreta e inteira é adotada: t e $t+1$ correspondem ao início de dois segmentos de tempo consecutivos, e o tamanho de cada segmento é uma constante, ϵ. Assume-se que as transmissões sem sucesso são devidas apenas a colisões (canal perfeito), e que colisões acontecem com a mesma probabilidade p independente do número de tentativas de retransmissão.

Seja m, o estágio de *backoff* máximo, um valor tal que $W_{max} = 2^m W_{min}$, onde W_{min} é a mínima janela de *backoff*, e será adotado a notação $W_i = 2^i W_{min}$, onde $i \in (0, m)$ é chamado estágio de *backoff*. Seja $S(t)$ o processo estocástico representando o estágio de *backoff* $(0, ..., m+R)$ da estação no tempo t, onde R representa a nova tentativa no estágio de *backoff* máximo. De $S(t)$ e $B(t)$, é possível estabelecer o modelo da cadeia de Markov bi-dimensional para o algoritmo de *backoff* mostrado na figura 4.2.

A figura 4.2 representa um processo bi-dimensional $\{S(t), B(t)\}$ na forma de uma cadeia de Markov de tempo discreto. Nesta figura, a unidade do período de *backoff* é expressa por W e é do mesmo tamanho de um segmento de tempo. p representa a probabilidade de colisão, que é um evento de probabilidade independente e com valor constante. Além disso, p expressa a probabilidade de colisão de uma mensagem de requisição de largura de banda durante a transmissão no canal de *upstream*.

Uma vez que se assume independência, e p é suposto ser um valor constante, é

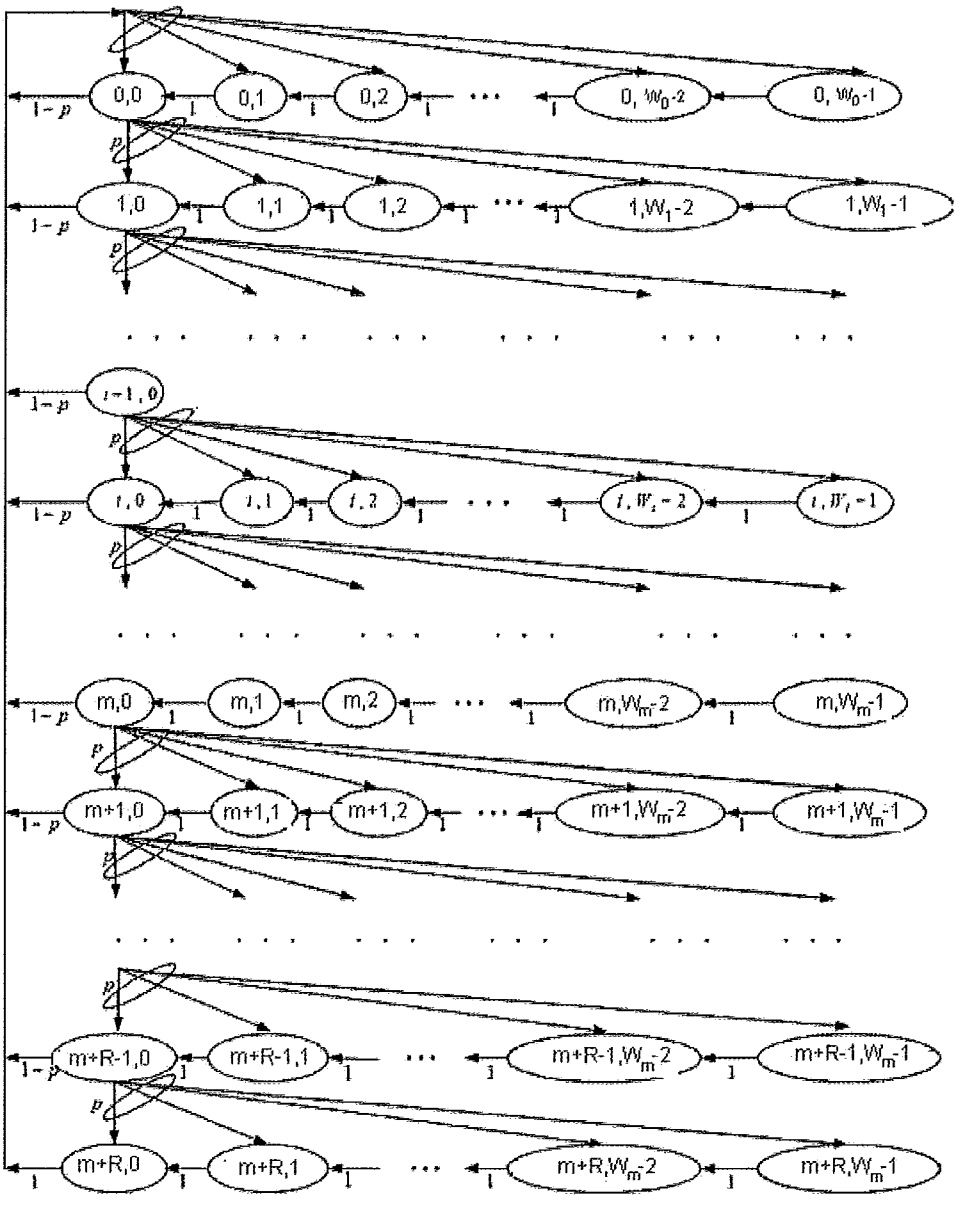


Figura 4.2: Modelo da cadeia de Markov para o algoritmo de *backoff*

possível modelar o processo bi-dimensional $\{S(t), B(t)\}$ como uma cadeia de Markov de tempo discreto como na figura 4.2. As seguintes notações são adotadas:

$$P_r\{i_1, k_1 | i_0, k_0\} = P_r\{s(t+1) = i_1, b(t+1) = k_1 | S(t) = i_0, B(t) = k_0\}$$

Serão discutidas agora as probabilidades de transição de estado não nulo desta cadeia de Markov.

- Quando $i < m$

  1. O contador de *backoff* decresce incondicionalmente no inicio de cada segmento de tempo.

  $$P_r\{i, k | i, k+1\} = 1 \qquad k \in (0, W_i - 2) \quad i \in (0, m)$$

  2. A estação entra no estado $\{0, k\}$ se ela verifica uma transmissão com sucesso e o contador de *backoff* é escolhido aleatoriamente entre $(0, W_0 - 1)$

  $$P_r\{0, k | i, 0\} = \frac{(1-p)}{W_0} \qquad k \in (0, W_0 - 1) \quad i \in (0, m)$$

  3. A estação escolhe um contador de *backoff* para o próximo estágio i depois de uma transmissão sem sucesso no estágio $i-1$ e o contador de *backoff* é escolhido aleatoriamente entre $(0, W_i - 1)$.

  $$P_r\{i, k | i-1, 0\} = \frac{p}{W_i} \qquad k \in (0, W_i - 1) \quad i \in (1, m)$$

- Quando $i > m$

  1. O contador de *backoff* decresce incondicionalmente no inicio de cada segmento de tempo.

  $$P_r\{i, k | i, k+1\} = 1 \qquad k \in (0, W_m - 2) \quad i \in (m+1, m+R)$$

2. A estação entra no estado $\{0, k\}$ se ela verifica uma transmissão com sucesso e o contador de *backoff* é escolhido aleatoriamente entre $(0, W_0 - 1)$.

$$P_r\{0, k|i, 0\} = \frac{(1-p)}{W_0} \qquad k \in (0, W_0 - 1) \quad i \in (m+1, m+R-1)$$

3. A estação escolhe um contador de *backoff* para o próximo estágio i depois de uma transmissão sem sucesso no estágio $i-1$. Quando $m \leq i \leq m+R$, o estágio de *backoff* não é acrescido, e permanece igual a m. Então, o contador de *backoff* é escolhido aleatoriamente entre $(0, W_m - 1)$.

$$P_r\{i, k|i-1, 0\} = \frac{p}{W_i} \qquad k \in (0, W_i - 1) \quad i \in (m+1, m+R)$$

4. A estação alcança o estágio final do procedimento de *backoff*. Se falha, o contador de *backoff* é zerado. Se ocorre um sucesso, o contador é reiniciado.

$$P_r\{0, k|m+R, 0\} = \frac{1}{W_0} \qquad k \in (0, W_0 - 1)$$

Sabe-se que o modelo é irredutível (i.e. cada estado pode ser alcançado de qualquer outro), aperiódico (i.e. um estado não pode ser alcançado dele mesmo) e recorrente não nulo. Assim, uma distribuição estacionária do modelo recorrente existe. Das 7 probabilidades de transição de estado único definidas acima, a distribuição estacionária, $b_{i,k}$, pode ser expressa como:

$$\begin{aligned} \text{If} \quad i \leq m, \quad b_{i,k} &= \lim_{t\to\infty} P_r\{S(t) = i, B(t) = k\} \qquad i \in (0, m) \quad k \in (0, W_i - 1) \qquad (4.1) \\ \text{If} \quad i > m, \quad b_{i,k} &= \lim_{t\to\infty} P_r\{S(t) = i, B(t) = k\} \qquad i \in (m+1, m+R) \quad k \in (0, W_m - 1) \end{aligned}$$

onde:

$$b_{i-1,0} \cdot p = b_{i,0} \rightarrow b_{i,0} = p^i \cdot b_{0,0} \qquad (4.2)$$

Devido a regularidade da cadeia, uma solução de forma fechada para esta cadeia de Markov pode ser obtida usando o seguinte procedimento. Primeiro, $b_{i,k}$ pode ser reescrito da forma abaixo:

$$b_{i,k} = \begin{cases} \frac{W_i - k}{W_i} \begin{cases} (1-p)\sum_{l=0}^{m+R-1} b_{l,0} + b_{m+R,0} & i = 0 \quad k \in (0, W_i - 1) \\ p \cdot b_{i-1,0} & 0 < i \leq m \quad k \in (0, W_i - 1) \end{cases} \\ \frac{W_m - k}{W_m} \cdot p \cdot b_{i-1,0} & m < i \leq m + R \quad k \in (0, W_m - 1) \end{cases} \tag{4.3}$$

Usando 4.2, 4.3 pode ser reescrito como:

$$\begin{cases} b_{i,k} = \frac{W_i - k}{W_i} b_{i,0} & i \in (0, m) \quad k \in (0, W_i - 1) \\ b_{i,k} = \frac{W_m - k}{W_m} b_{i,0} & i \in (m+1, m+R) \quad k \in (0, W_m - 1) \end{cases} \tag{4.4}$$

De 4.2 e 4.4, segue que os valores de $b_{i,k}$ podem ser expressos como funções do valor de $b_{0,0}$ e a probabilidade condicional de colisão, p. Finalmente, $b_{0,0}$ é determinada usando a lei da conservação da probabilidade, i.e.

$$\begin{aligned} 1 &= \sum_{i=0}^{m} \sum_{k=0}^{W_i - 1} b_{i,k} + \sum_{i=m+1}^{m+R} \sum_{k=0}^{W_m - 1} b_{i,k} \\ &= \sum_{i=0}^{m} b_{i,0} \frac{W_i + 1}{2} + \sum_{i=m+1}^{m+R} b_{i,0} \frac{W_m + 1}{2} \\ &= \frac{b_{0,0}}{2} \frac{W_{min}[(1-p)(1-(2p)^{m+1}) + (1-2p)2^m p^{m+1}(1-p^R)] + (1-2p)(1-p^{m+R+1})}{(1-2p)(1-p)} \end{aligned} \tag{4.5}$$

onde

$$b_{0,0} = \frac{2(1-2p)(1-p)}{W_{min}[(1-p)(1-(2p)^{m+1}) + (1-2p)2^m p^{m+1}(1-p^R)] + (1-2p)(1-p^{m+R+1})} \tag{4.6}$$

Dados os valores de R, W_{min} e p, a probabilidade de estado fixo do modelo pode ser calculada de 4.2 até 4.6.

Seja τ a probabilidade de estado fixo de uma estação transmitir durante qualquer segmento de tempo. Na rede, uma estação apenas transmite quando seu contador

de *backoff* é igual a zero (i.e. uma estação transmite para qualquer i de $b_{i,0}$). Temos então:

$$\tau = \sum_{i=0}^{m+R} b_{i,0} \tag{4.7}$$
$$= \frac{2(1-2p)(1-p^{m+R+1})}{W_{min}[(1-p)(1-(2p)^{m+1}) + (1-2p)\cdot 2^m \cdot p^{m+1}(1-p^R)] + (1-2p)(1-p^{m+R+1})}$$

Uma colisão ocorre quando duas ou mais estações transmitem durante o mesmo segmento de tempo. Assim, a probabilidade de colisão, p, de uma mensagem de requisição de largura de banda é dada por:

$$p = 1 - (1-\tau)^{N-1} \tag{4.8}$$

As equações 4.7 e 4.8 representam um sistema não linear com dois parâmetros não conhecidos, τ e p. Substituindo 4.8 em 4.7 obtém-se uma equação com apenas um parâmetro não conhecido, τ. Resolvendo essa equação para τ, tem-se a probabilidade p, o que permite a subseqüente derivação da distribuição estacionária substituindo $b_{0,0}$ e p em 4.4. Logo:

$$b_{i,k} = \frac{2^i W_{min} - k}{2^i W_{min}} p^i \frac{2(1-2p)(1-p)}{W_{min}[(1-p)(1-(2p)^{m+1}) + (1-2p)2^m p^{m+1}(1-p^R)] + (1-2p)(1-p^{m+} \tag{4.9}$$

Ainda é necessário definir alguns parâmetros para o cálculo do retardo do mecanismo de reserva. Seja P_{tr} a probabilidade de transmissão no sistema durante qualquer segmento de tempo. Em outras palavras, P_{tr} denota a probabilidade de pelo menos uma estação transmitir durante um segmento escolhido aleatoriamente. Note que esse valor difere de τ, que indica a probabilidade de uma estação particular transmitir durante um segmento de tempo aleatório. Desde que N estações disputam o canal e transmitem com probabilidade τ, mostra-se então que:

$$P_{tr} = 1 - (1-\tau)^N \tag{4.10}$$

Além disso, a probabilidade P_s é definida como a probabilidade de transmissão com sucesso, que representa o caso de exatamente uma estação transmitir no canal, condicionada pela probabilidade P_{tr}, isto é:

$$P_s = \frac{\binom{N}{1}\tau(1-\tau)^{N-1}}{P_{tr}} = \frac{N\tau(1-\tau)^{N-1}}{1-(1-\tau)^N} \tag{4.11}$$

4.2.3 Análise do Retardo Médio das Mensagens de Requisição de Largura de Banda

Na seção anterior, foi encontrada a probabilidade de uma estação enviar uma mensagem de requisição de largura de banda em um segmento de contenção aleatório (isto é, τ), e a probabilidade de colisão de uma mensagem de requisição transmitida no canal (isto é, p). Nesta seção, será calculado o retardo de acesso das mensagens de requisição de largura de banda.

Como dito antes, o retardo de acesso é o intervalo entre o tempo da chegada da mensagem na SS e o tempo quando ela é enviada pela SS. Este tempo pode ser dividido em duas partes: o retardo de acesso da requisição (tempo entre a chegada da mensagem e a transmissão com sucesso da requisição) e o retardo de escalonamento da requisição (tempo entre a transmissão com sucesso da requisição e o início da transmissão do dado). Esses dois intervalos são não sobrepostos. Novamente, nesta seção, calcula-se o retardo de reserva, ou seja, o retardo de acesso da mensagem de requisição de largura de banda.

Define-se D_r, o retardo médio das mensagens de requisição de largura de banda, como o tempo gasto entre sua geração e sua recepção com sucesso. Colisões podem ocorrer durante o processo de transmissão até ocorrer o sucesso. Assim, tem-se:

$$E[D_r] = E[N_c](E[\delta] + T_c) + (E[\delta] + T_s) \tag{4.12}$$

onde $E[N_c]$ é o valor esperado do número de colisões experimentado por uma mensagem de requisição de largura de banda antes da recepção com sucesso pela BS, $E[\delta]$ é o retardo médio do contador de *backoff* especificado por uma estação antes de acessar o canal em condições ocupadas, T_c é o tempo de duração de uma colisão e finalmente T_s é o tempo gasto para uma transmissão com sucesso.

Do comportamento de uma transmissão (isto é, colide continuamente antes da recepção com sucesso) e da definição de valor médio, temos que a variável aleatória N_c comporta-se como uma distribuição geométrica de parâmetro P_s. Assim, o valor médio de N_c é dado por:

$$E[N_c] = \sum_{i=1}^{\infty} i(1-P_s)^i P_s = P_s(1-P_s)\frac{\partial}{\partial P_s}\frac{1-P_s}{P_s} \tag{4.13}$$

Logo,

$$E[N_c] = \frac{1}{P_s} - 1 \tag{4.14}$$

Será analisado agora o retardo devido ao contador de *backoff*, que depende do valor do contador e da duração em que ele permanece congelado quando inicia o período de transmissão de dados, garantido as estações pela BS. Quando o contador de uma estação está no estado $b_{i,k}$, um intervalo de tempo de k segmentos é requerido para o contador alcançar o estado $b_{i,0}$. Este intervalo é denotado pela variável aleatória β, cujo valor médio é dado por:

$$\begin{aligned} E[\beta] = & \sum_{i=0}^{m}\sum_{k=1}^{W_i} k b_{i,k} + \sum_{i=m+1}^{m+R}\sum_{k=1}^{W_{m-1}} k b_{i,k} \qquad (4.15) \\ = & \frac{b_{0,0}}{6}\left\{\frac{W_{min}^2[(1-p)(1-(4p)^{m+1})+4^m(1-4p)p^{m+1}(1-p^R)]-(1-4p)(1-p^{m+1})+p^{m+1}(1-p^R)}{(1-4p)(1-p)}\right\} \end{aligned}$$

O tempo pelo qual o contador da estação permanece congelado é denotado por Φ. Quando o contador congela, ele permanece inativo pela duração de um período reservado para a transmissão de dados. Para calcular o tempo $E[\Phi]$ pelo qual o

contador permanece inativo, é necessário estabelecer $E[N_q]$, isto é, o número médio de vezes que uma estação deve esperar pela oportunidade de transmissão de outras estações antes do seu contador alcançar 0. $E[N_q]$ é baseado em $E[\beta]$, isto é, o retardo médio de *backoff* de cada estação, e em μ, isto é, o tamanho do período de reserva. Então, pode ser mostrado que

$$E[N_q] = E[\beta]/\mu$$

Além disso:

$$E[\Phi] = E[N_q](E[T_q] - \mu) \tag{4.16}$$

Onde, $E[T_q]$ é definido como o tempo médio de duração de um quadro do 802.16. Considerando μ um parâmetro constante.

De 4.15 e 4.16, pode ser mostrado que:

$$E[\delta] = E[\beta] + E[\Phi] \tag{4.17}$$

Substituindo 4.14 e 4.17 em 4.12 pode-se calcular o retardo médio do período de reserva, ou seja, das mensagens de requisição de largura de banda. Note que os intervalos de tempo e relações acima são medidas na mesma unidade.

4.3 Retardo na Fase de Alocação de Dados

Esta seção propõe um modelo para alocação de dados com duas classes de prioridade de tráfego que suporta serviço de tempo real e serviço de tempo não-real. O tráfego de tempo real tem prioridade com interrupção sobre o tráfego de tempo não-real.

4.3.1 Suposições e Definições

Assume-se que há duas classes de tráfego com diferentes requerimentos de serviço, tráfego de tempo real transmitido por UGS e tráfego de tempo não-real transmitido através de mecanismo de reserva ou *piggybacking*. A largura de banda de um UGS é predeterminada, a BS conhece seu tamanho e seu tempo de serviço. Assume-se que há requisições virtuais para tráfego de tempo real, cujo tamanho requisitado é igual ao tamanho da garantia não solicitada UGS e o tempo de chegada de requisições virtuais de tráfego de tempo real é igual ao tempo de garantia nominal UGS. A BS tem um *buffer* virtual para acomodar requisições virtuais do tráfego de tempo real e um *buffer* para acomodar requisições do tráfego de tempo não-real.

A chegada das requisições virtuais do tráfego de tempo real e requisições do tráfego de tempo não-real seguem um processo de Poisson independente com taxas de chegada λ_1 e λ_2, respectivamente. O tempo de serviço das requisições virtuais do tráfego de tempo real e das requisições do tráfego de tempo não-real é assumido ser independente e identicamente distribuído com distribuição geral. Seja ν_1 e ν_2 o tempo de serviço médio das requisições virtuais de tempo real e das requisições do tráfego de tempo não-real, respectivamente.

Como foi visto, o canal de *upstream* é composto de um *stream* de segmentos, incluindo segmentos de dados e segmentos de contenção. Não será considerado tempo de guarda. A disciplina de serviço para ambos os tipos de tráfego é FCFS (First-Come-First-Served). Considera-se que μ segmentos são alocados para oportunidades de transmissão de reserva em cada MAP.

Como mencionado, a largura de banda para o tráfego de tempo real é predeterminada portanto, não é necessário enviar requisição de largura de banda para ele. Define-se como o retardo do tráfego de tempo real como o intervalo de tempo entre o tempo de garantia atual e o tempo de garantia nominal. Após chegar a BS, uma requisição do tráfego de tempo não-real deverá esperar no *buffer* até o escalonador da BS alocar largura de banda para ele no próximo MAP. O retardo do tráfego de tempo não-real é definido como o tempo entre quando uma requisição

para o tráfego de tempo não-real chega na BS e quando o último bit deste pacote de dados requisitado chega na BS.

4.3.2 Modelo para Análise

A relação de tempo definida pelo MAP é mostrado na figura 4.3. Assume-se que o tempo definido pelo último MAP inicia-se no tempo t_1 e termina no tempo t_2. A BS recebe requisições de tempo não-real de t_1 até t_2. No tempo t_2, a BS aloca segmentos para as requisições virtuais do tráfego de tempo real, requisições do tráfego de tempo não-real e segmentos de contenção para formar um MAP. Considerando que o tempo de guarda e o retardo de processamento da BS e SS são ignorados, o tempo de segmentos definido pelo MAP é de t_2 até t_3. De acordo com as suposições da seção 4.3.1, um modelo que é uma variante do modelo *Leaky Bucket bufferizado* com prioridade [34] é criado como mostrado na figura 4.4. Há um *token pool* no modelo. O tamanho do *token pool* é 4096, cada *token* tem duração de um segmento. Inicialmente, o *token pool* está completo com 4096 *tokens*. Há um *buffer* virtual para acomodar as requisições que chegam do tráfego de tempo real, um *buffer* para acomodar as requisições do tráfego de tempo não-real e outro *buffer* virtual para acomodar μ segmentos de contenção. O número de *tokens* no *token pool* é decrementado de uma unidade para cada segmento alocado. De acordo com o mecanismo de requisição de largura de banda, requisições do tráfego de tempo não-real que chegam de t_1 até t_2 serão servidas no próximo MAP, logo o início de tempo do *buffer* para o tráfego de tempo não-real é t_2. Requisições virtuais de tempo real chegando durante o tempo definido pelo próximo MAP serão servidos no próximo MAP, logo o início do tempo do *buffer* de tempo real é iniciado no tempo t_3.

O modelo trabalha da seguinte forma:

1. A estação recebe requisições para o tráfego de tempo não-real de t_1 até t_2 e coloca-as no *buffer* de tráfego de tempo não-real na ordem de chegada;
2. A BS inicia a alocação de segmentos para formar um MAP em t_2. A BS aloca segmentos para as requisições do tráfego de tempo não-real armazenadas no

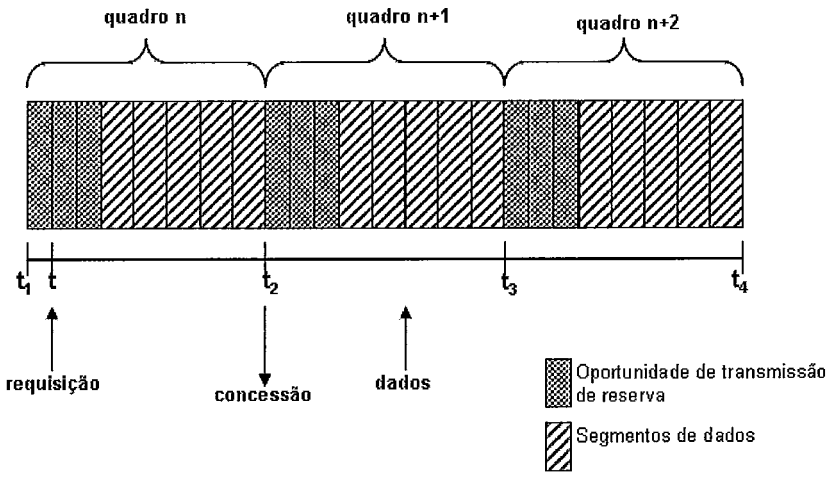


Figura 4.3: Relação do Tempo no MAP

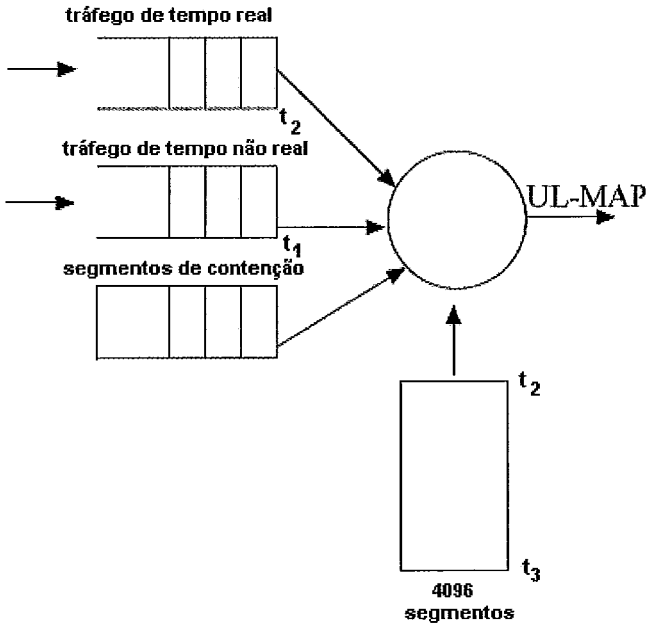


Figura 4.4: O modelo analítico

buffer quando o *buffer* de tempo real está vazio e pára a atribuição com a chegada de requisições virtuais do tráfego de tempo real; ele volta a atribuição das requisições do tráfego de tempo não-real quando ele termina as atribuições do tráfego de tempo real e o *buffer* virtual do tráfego de tempo real está vazio;

3. Quando o *buffer* do tráfego de tempo não-real está vazio, a BS inicia a atribuição de μ segmentos, e pára com a chegada de requisições virtuais para tráfego de tempo real.

4. Depois de μ segmentos de contenção terem sido atribuídos até t_3 a alocação termina. A estação reinicia o *token pool* com os próximos 4096 *tokens*, cujo primeiro *token* representa o tempo t_4. Ignorando o retardo no processo, a operação de (2) até (4) pode ser considerada realizada imediatamente;

5. O procedimento de (1) até (4) é repetido.

4.3.3 Retardo Médio para o Tráfego de Tempo Real

De acordo com as suposições da seção 4.3.1 e o modelo analítico da seção 4.3.2, a alocação de segmentos para requisições para o tráfego de tempo não-real e segmentos de contenção será interrompida com a chegada de requisições virtuais de tempo real. Visto que as requisições virtuais de tempo real tem prioridade com interrupção, o processo $\{x_t^{(1)}, t \in [0, \infty)\}$, onde $x_t^{(1)}$, o número de requisições virtuais presentes no sistema no tempo t, é o processo de fila M/G/1. Todas as probabilidades relacionadas a $x_t^{(1)}$ podem ser imediatamente obtidas. Portanto, o retardo médio para o tráfego de tempo real pode ser obtido por [9]:

$$E[D_1] = \frac{\lambda_1 E[V^2]}{2(1 - \lambda_1 E[V])} \tag{4.18}$$

onde λ_1 é a taxa de chegada de requisições virtuais do tráfego de tempo real e V é a variável aleatória do tempo de serviço das requisições virtuais do tráfego de tempo real. Logo, $E[D_1]$ é o retardo médio do tráfego de tempo real, ver apêndice A, equação A.11.

4.3.4 Retardo Médio do Tráfego de Dados de Tempo Não-real

Assume-se que o número de requisições para o tráfego de tempo não-real no *buffer* do tráfego de tempo não-real é $l-1$ e a l-ésima requisição para o tráfego de tempo não-real chega na BS no tempo $t' \in [t_1, t_2]$. As requisições para o tráfego de tempo não-real são servidas na ordem de chegada, durante o qual o serviço será interrompido e as requisições virtuais do tráfego de tempo real serão servidas, se elas chegarem. Assuma que g requisições virtuais para o tráfego de tempo real chegem antes da l-ésima requisição do tráfego de tempo não-real ser servida. De acordo com a disciplina de serviço e a relação de tempo da seção 4.3.2, o retardo da i-ésima requisição para tráfego de tempo não-real será:

$$D_2 = t_2 - t' + \sum_{i=1}^{l} S_i + \sum_{j=1}^{g} V_j \tag{4.19}$$

onde S_i é a variável aleatória que representa o tempo de serviço para a i-ésima requisição do tráfego de tempo não-real, V_j é a variável aleatória que representa o tempo de serviço para a j-ésima requisição virtual do tráfego de tempo real, t_2 é o tempo do início do próximo MAP e t' é o tempo de chegada da l-ésima requisição para o tráfego de tempo não-real. Então o retardo médio para o tráfego de tempo não-real é:

$$E[D_2] = E[T_{MAP}] - E[t'] + E[l]E[S] + E[g]E[V] \tag{4.20}$$

onde $E[T_{MAP}]$ é o tempo médio definido pelo MAP, $E[S]$, $E[V]$, são o tempo de serviço médio para os tráfego de tempo não-real e de tempo real, respectivamente. $E[t']$ é o tempo médio de chegada de uma requisição para tráfego de tempo não-real no intervalo $[t_1, t_2]$, $E[l]$ é o número médio de chegadas de requisições para o tráfego de tempo não-real de 0 até $E[t']$ e $E[g]$ é o número médio de requisições virtuais para o tráfego de tempo real que chegam antes da $i-$ésima requisição para o tráfego de tempo não-real ser servida.

Em condições estacionárias, o tempo definido pelo próximo MAP consiste do tempo de serviço das requisições virtuais para o tráfego de tempo real chegando durante o tempo do próximo MAP, o tempo de serviço das requisições para o tráfego de tempo não-real chegando de t_1 até t_2 e o tempo ocupado por μ segmentos de contenção. Que é,

$$\mu + \sum_{i=1}^{G} V_i + \sum_{j=1}^{L} S_j = T_{MAP} \tag{4.21}$$

onde T_{MAP} é o tempo definido pelo próximo MAP, G é a variável aleatória que representa o número de requisições virtuais para o tráfego de tempo real chegando durante o tempo do próximo MAP e L é a variável aleatória que representa o número de requisições para o tráfego de tempo não-real chegando de t_1 até t_2, ver apêndice A, equação A.12. Aqui o tempo é medido em unidade de número de segmentos. Pegando a média de 4.21 e substituindo $E[G] = \lambda_1 E[T_{MAP}]$, $E[L] = \lambda_2 E[T_{MAP}]$ tem-se,

$$E[T_{MAP}] = \frac{\mu}{1 - (\lambda_1 E[V] + \lambda_2 E[S])} \tag{4.22}$$

A chegada de requisições para o tráfego de tempo não-real comporta-se como um processo de Poisson; t' é igualmente distribuído, e pode-se derivar [9]:

$$E[g] = \frac{\lambda_1 \lambda_2 E[T_{MAP}] E[S]}{2(1 - \lambda_1 E[V])} \tag{4.23}$$

$$E[l] = \lambda_2 E[T_{MAP}]/2 \tag{4.24}$$

$$E[t'] = E[T_{MAP}]/2 \tag{4.25}$$

Substituindo 4.22 até 4.25 em 4.20, obtém-se o retardo médio para tráfego de tempo não-real, ver apêndice A, equação A.13

$$E[D_2] = \frac{\mu}{2(1-(\lambda_1 E[V] + \lambda_2 E[S]))}\left(1 + \lambda_2 E[S] + \frac{\lambda_1 \lambda_2 E[S]E[V]}{(1-\lambda_1 E[V]}\right) \quad (4.26)$$

se $\rho_1 = \lambda_1 E[V] = \lambda_1 \nu_1$ é a carga do tráfego de tempo real oferecida ao sistema, isto é, o número médio de mensagens que chegam no sistema, geradas pelas G estações com fluxo de tempo real, durante o tempo médio de transmissão de uma mensagem, $\rho_2 = \lambda_2 E[S] = \lambda_2 \nu_2$ é a carga do tráfego de tempo não-real oferecida ao sistema, isto é, o número médio de mensagens que chegam no sistema, geradas pelas L estações com fluxo de tempo não-real, durante o tempo médio de transmissão de uma mensagem, então $\rho = \rho_1 + \rho_2$ é a carga total oferecida ao sistema.

4.4 Retardo Total

Pelo modelo construído nas seções anteriores, pode-se encontrar o retardo médio total para o tráfego de tempo real por:

$$E[D_1]_T = \frac{\lambda_1 E[V^2]}{2(1-\lambda_1 E[V])} \quad (4.27)$$

O retardo médio total para o tráfego de tempo não-real é dado pela soma dos componentes do retardo de acesso devido a camada MAC e a alocação dos dados, que também é influenciada pelo tráfego mais prioritário, ou seja:

E$[D_2]_T$ = retardo na fase de requisição de largura de banda + retardo na fase de alocação de dados

$$E[D_2]_T = E[D_r] + E[D_2]$$

$$E[D_2]_T = E[N_f](E[BD] + T_f) + (E[BD] + T_s) + \frac{\mu}{2(1-(\lambda_1 E[V] + \lambda_2 E[S]))}\left(1 + \lambda_2 E[S] + \frac{\lambda_1 \lambda_2 E[S]E[V]}{(1-\lambda_1 E[V])}\right) \quad (4.28)$$

4.5 Considerações Finais

Neste capítulo foi apresentada a modelagem e a análise de desempenho do padrão IEEE 802.16. Este modelo permite o cálculo de métricas importantes de desempenho, tais como tamanho da fila e atraso das mensagens de requisição bem como das mensagens de dados. Os resultados podem ser facilmente obtidos a partir das equações que modelam o sistema. A aplicação da teoria demonstrada neste capítulo é baseada na teoria de filas e cadeia de Markov.

Através da modelagem analítica apresentada neste capítulo, alguns resultados numéricos para o retardo total das mensagens serão apresentados no próximo capítulo, para ilustrar o desempenho do protocolo da camada MAC do padrão IEEE 802.16.

Capítulo 5

Resultados Obtidos

PARA analisar o comportamento do protocolo MAC do padrão IEEE 802.16 com relação ao retardo provocado pelo mecanismo de escalonamento do *uplink*, utilizado para as diferentes classes de tráfego, este capítulo apresenta alguns resultados numéricos obtidos com o modelo analítico proposto no capítulo anterior. O retardo das mensagens de tempo real e de tempo não-real são calculados em dois cenários distintos, onde pode-se comparar qual é a influência de uma alta carga de fluxos de diferentes prioridades. Além disso, investiga-se o mecanismo de requisição das mensagens de largura de banda, para diferentes parâmetros de *backoff* e vários tamanhos do período de contenção. Por fim, através da ferramenta de simulação NS-2 (*Network Simulator*) [35], foi avaliado o modelo analítico proposto.

5.1 Análise da Fase de Alocação de Dados

Para avaliar o nível de diferenciação obtido com o modelo descrito no capítulo anterior para a fase de alocação de dados, serão considerados dois cenários distintos, onde em cada cenário existe uma porcentagem diferente para cada tipo de tráfego gerado pelas estações, como mostra a Tabela 5.1. Estas classes podem ser mapeadas para os quatro tipos de serviços oferecidos pelo padrão IEEE 802.16 da seguinte maneira: a classe de tempo real (que é a mais prioritária) representa os serviços UGS e rtPS e a classe de tempo não-real representa os serviços nrtPS e BE. A diferença entre os cenários é que, no Cenário I existe um maior número de estações transmitindo tráfego de tempo real, 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. As fórmulas fechadas derivadas no capítulo anterior, para a modelagem proposta, foram implementadas no software MATLAB e serão utilizadas para a geração dos resultados apresentados neste capítulo.

Classes de tráfego	**Cenário I**	**Cenário II**
Tempo real	70%	30%
Tempo não-real	30%	70%

Tabela 5.1: Cenários de tráfego utilizados na modelagem analítica.

Os retardos médios do tráfego de tempo real e não-real são examinados, nos cenários I e II, como função da intensidade de tráfego $\rho_1 = \lambda_1\nu_1$ e $\rho_2 = \lambda_2\nu_2$, onde $\rho = \rho_1 + \rho_2$, como foi mostrado nas equações 4.18 e 4.26 da seção 4.3.3. As diferentes cargas, nos cenários I e II, são reportadas na tabela 5.2. Os valores dos parâmetros usados para encontrar o retardo médio dos tráfegos de tempo real e não-real são mostrados na tabela 5.3, considerando, o tamanho do período de contenção, $\mu = 8$ segmentos.

As Figuras 5.1 e 5.2 ilustram o tempo médio de espera na fila para cada classe de prioridade em relação ao tráfego oferecido no canal. Na figura 5.1, o eixo y

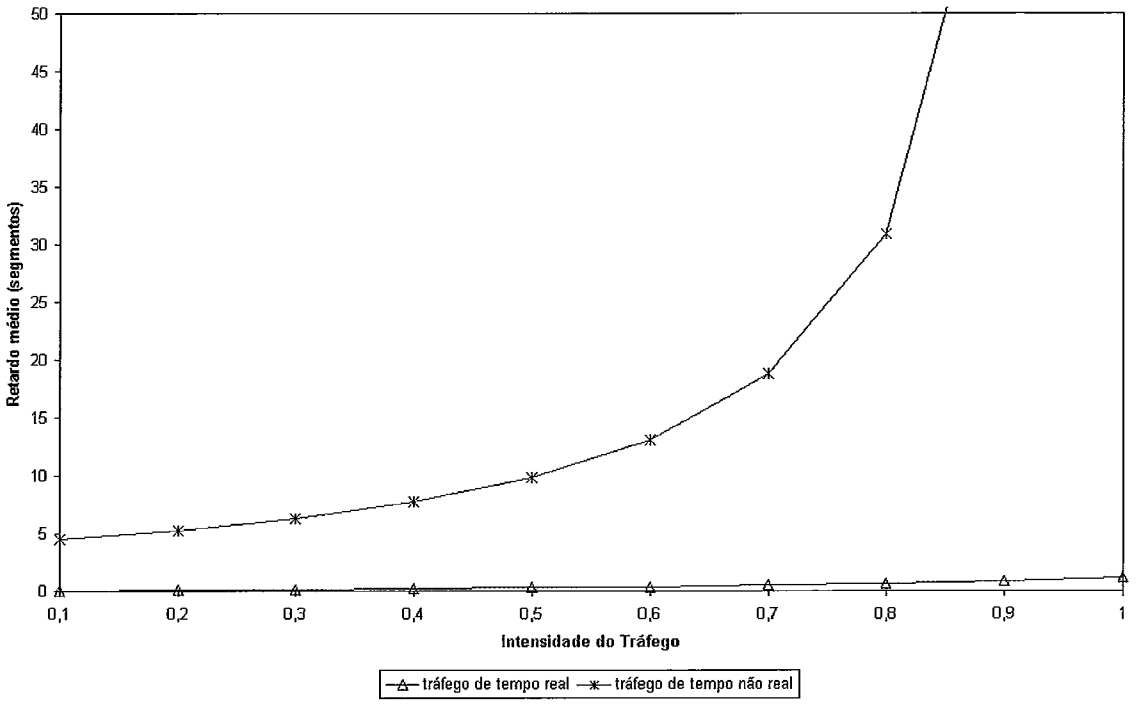


Figura 5.1: Retardo médio do tráfego de tempo real e de tempo não-real sob o cenário I.

representa o retardo médio para as mensagens de dados de tempo real e não-real sob o cenário I, onde há a ocorrência de maior fluxo de tráfego de tempo real. E o eixo x mostra a carga total agregada dos tráfegos de tempo real e não-real, onde, por exemplo, para a carga de 0,1, tem-se 0,07 de carga de tráfego real e 0,03 de carga de tráfego não-real. Na figura 5.2, o eixo y representa o retardo médio para as mensagens de dados de tempo real e não-real considerando o cenário II no qual, existe maior fluxo do tráfego de tempo não-real. E o eixo x mostra a carga total

Carga oferecida	**Cenário I**	**Cenário II**
Tempo real	0,07; 0,14; 0,21; 0,28; 0,35; 0,42; 0,49; 0,56; 0,63; 0,70	0,03; 0,06; 0,09; 0,12; 0,15; 0,18; 0,21; 0,24; 0,27; 0,30
Tempo não-real	0,03; 0,06; 0,09; 0,12; 0,15; 0,18; 0,21; 0,24; 0,27; 0,3	0,07; 0,14; 0,21; 0,28; 0,35; 0,42; 0,49; 0,56; 0,63; 0,7

Tabela 5.2: Intensidade de Tráfego.

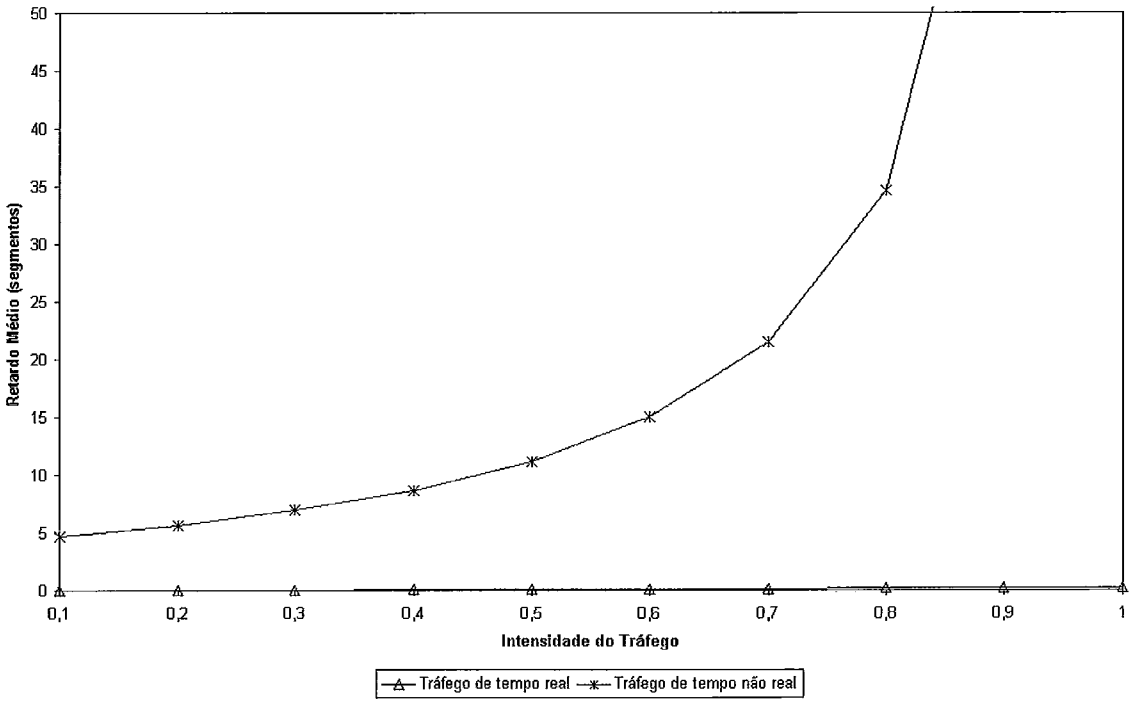


Figura 5.2: Retardo médio do tráfego de tempo real e de tempo não-real sob o cenário II.

agregada dos tráfegos de tempo real e não-real, onde, por exemplo, para a carga de 0,1, tem-se 0,03 de carga do tráfego real e 0,07 de carga do tráfego não-real.

Pelas figuras 5.1 e 5.2, observa-se que, o tempo de espera para o tráfego de alta prioridade (classe de tempo real) é bem menor em relação ao de menor prioridade, até mesmo no Cenário II onde existe uma maior probabilidade dos tráfegos de baixa prioridade. O modelo consegue diferenciar eficientemente as classes de tráfego, garantindo menor tempo de espera na fila para as mensagens de maior prioridade. Com isso, pode-se demonstrar a conformidade da modelagem das mensagens de dados com o padrão IEEE 802.16.

5.2 Análise da Requisição de Largura de Banda

Nesta seção, investiga-se a relativa efetividade do mecanismo de largura de banda com o tráfego de dados. Os valores dos parâmetros em consideração são reportados

na tabela 5.3. Em todos os cenários dessa seção foi fixado a carga do tráfego de tempo não-real, em $\rho_2 = 0.5$, para que os resultados não sejam influenciados pela variação da carga de dados. Visto que o serviço de escalonamento BE é empregado, as estações requisitam largura de banda para a BS enviando requisições de largura de banda via contenção.

A figura 5.3 avalia o impacto da janela de contenção inicial de *backoff* no desempenho, em termos do retardo médio das mensagens de requisição de largura de banda, considerando os parâmetros $\{m = 6, R = 10, \mu = 8\}$. É mostrado o retardo médio com N = 8, 16, 32, 64, 128, 256 e 512 estações, quando a janela mínima de *backoff* aumenta de 4 até 64. Sem levar em consideração o número de estações, ou seja, mantendo a quantidade de estações que requisitam largura de banda constante, e aumentando o tamanho da janela inicial de *backoff* o retardo médio das mensagens diminui até um valor mínimo, quando volta a aumentar. Esse comportamento pode ser explicado pelos gráficos 5.4 e 5.5, assim, quando o número de estações, disputando o meio, é grande e a janela inicial é pequena ocorrem muitas colisões afetando o retardo das mensagens. Por outro lado, maiores valores da janela reduzem a probabilidade de colisão entre as requisições de largura de banda no inicio do processo de contenção, fazendo com que o retardo diminua. Porém, janelas iniciais grandes contribuem para o aumento do número de segmentos vazios, e portanto o retardo volta a aumentar, devido ao número de segmentos ociosos ocasionados pelo maior adiamento causado pela maior janela inicial.

A figura 5.4 mostra a probabilidade de sucesso, que é definida como a probabilidade de transmissão com sucesso, que representa o caso de exatamente uma estação transmitir no canal condicionada pela probabilidade P_{tr}, ou seja, representa o sucesso dentre todas as mensagens de requisição transmitidas, considerando os parâmetros $\{m = 6, R = 10, \mu = 8\}$. Portanto, aumentar o tamanho da janela inicial, diminui as colisões das mensagens, para todos os valores de estações disputando o meio.

Para melhor avaliar o que ocorre no processo de disputa no canal, a figura 5.5 mostra a probabilidade de transmissão durante um segmento aleatório, considerando os parâmetros $\{m = 6, R = 10, \mu = 8\}$. Note que esse valor depende da janela

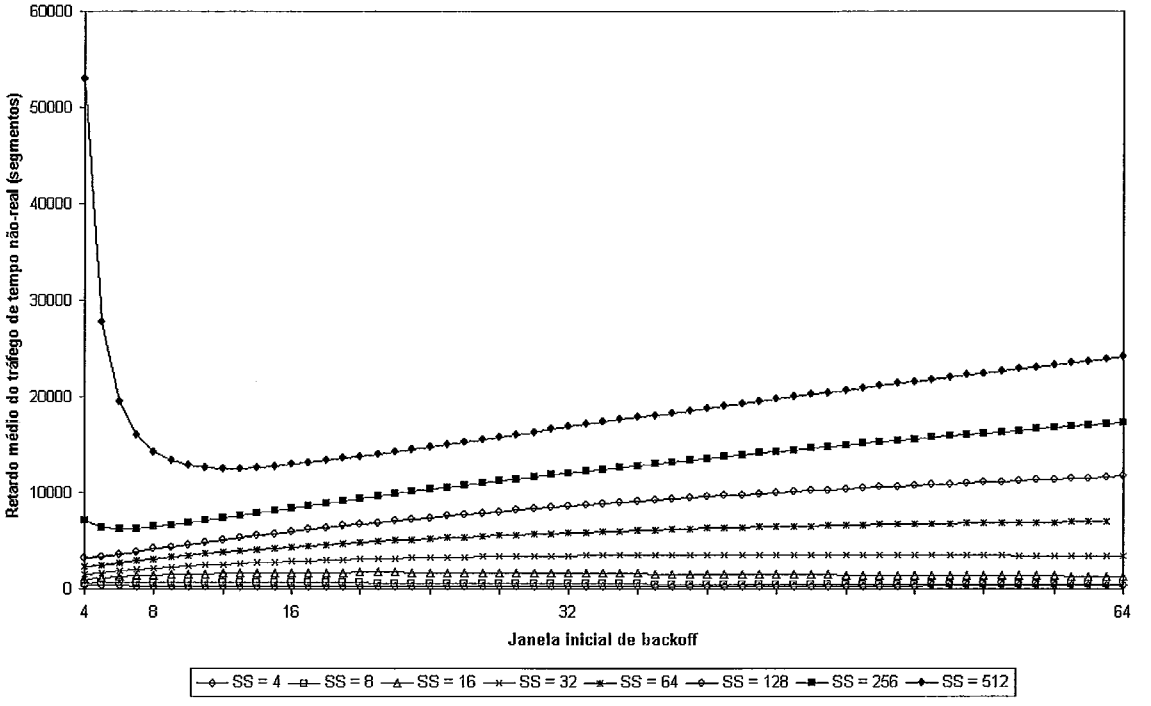


Figura 5.3: Retardo médio total do tráfego de tempo não-real versus a janela mínima de *backoff*.

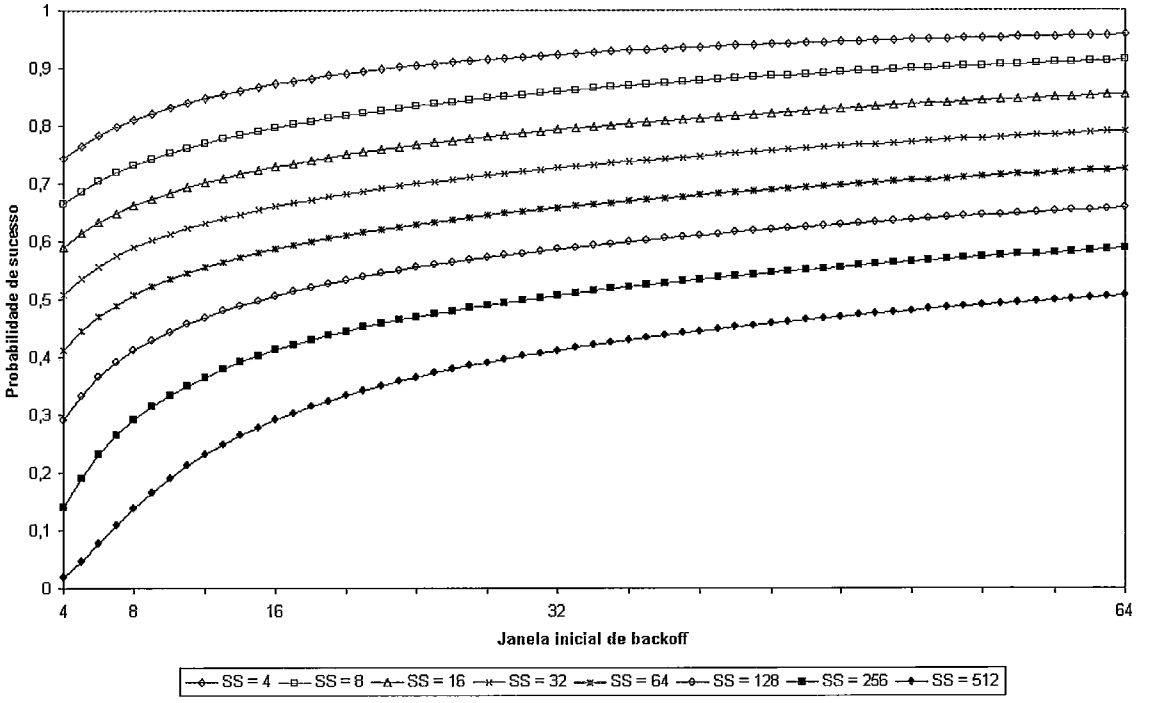


Figura 5.4: Probabilidade de transmissão com sucesso da mensagem de requisição de largura de banda versus a janela mínima de *backoff*.

Parâmetro	Valores
Janela inícial de *backoff* - W_{min}	8 segmentos
Janela de *backoff* máxima - W_{max}	64 segmentos
Número de segmentos de contenção - μ	4, 8, 16, 32
Média do tempo de serviço para as mensagens de tempo real - ν_1	1 segmento
Média do tempo de serviço para mensagens de tempo não-real - ν_2	1 segmento
Estágio máximo de *backoff* - m	3, 5, 7, 9, 11, 13, 15 e 16
Número de tentativas no estágio máximo de *backoff* - R	0, 1, 3, 5, 7, 9, 11 e 13
Número de Estações - N	4 - 512
Tamanho do quadro de *uplink*	50 segmentos

Tabela 5.3: Parâmetros do modelo.

inicial de *backoff* e do número de estações na rede. Portanto, para uma maior janela de contenção, a probabilidade da estação transmitir diminui, aumentando conseqüentemente o retardo da mensagem de requisição de largura de banda devido à espera pela oportunidade de transmissão da estação.

Na figura 5.6, o eixo x indica a quantidade de estações que disputam pelo acesso ao canal de *uplink* no período de contenção, enviando mensagens de requisição de largura de banda, e o eixo y indica o retardo total das mensagens de tempo não-real para diferentes valores da janela inicial de *backoff*, considerando os parâmetros $\{m = 6, R = 10, \mu = 8\}$. Quando o número de estações aumenta, de 4 até 512, o retardo também cresce devido ao aumento do número de colisões experimentadas pelas mensagens de requisição de largura de banda. Além disso, o retardo do sistema é afetado pelo tamanho da janela inicial de *backoff*. Pequenas janelas iniciais de *backoff* fornecem retardos menores até 220 estações, quando o retardo cresce rapidamente. Esse comportamento pode ser explicado pelo fato de que mesmo as mensagens de requisição colidindo mais vezes, o período de adiamento para nova

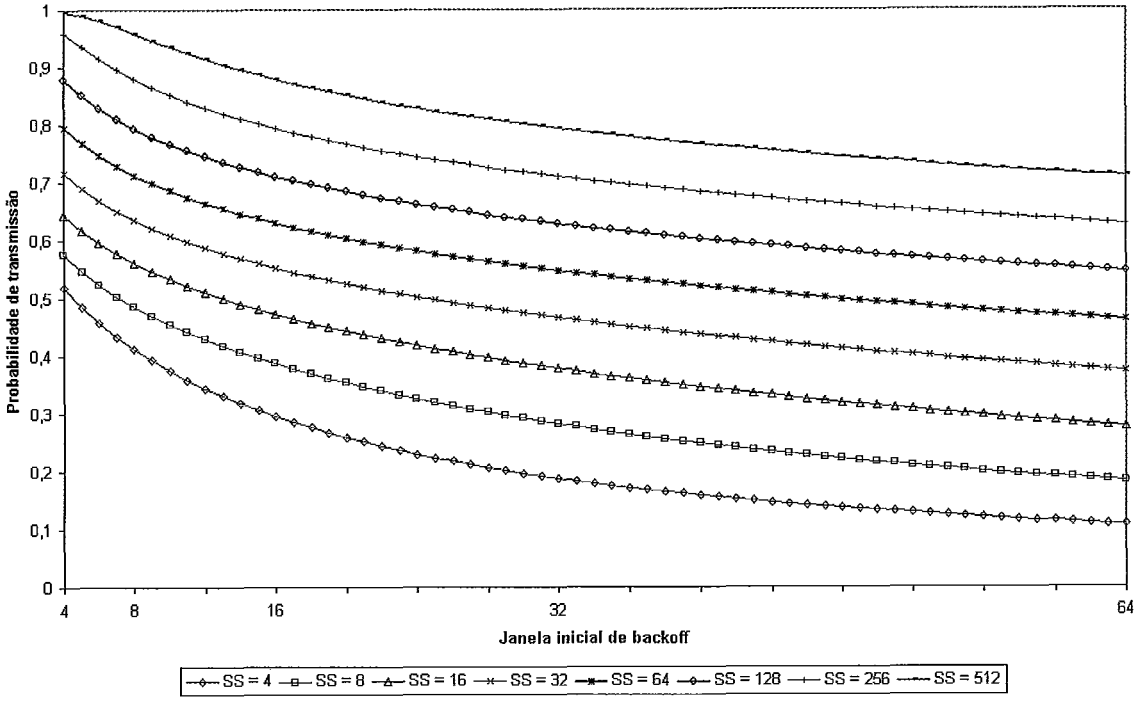


Figura 5.5: Probabilidade de transmissão de uma mensagem de requisição de largura de banda versus a janela inicial de *backoff*.

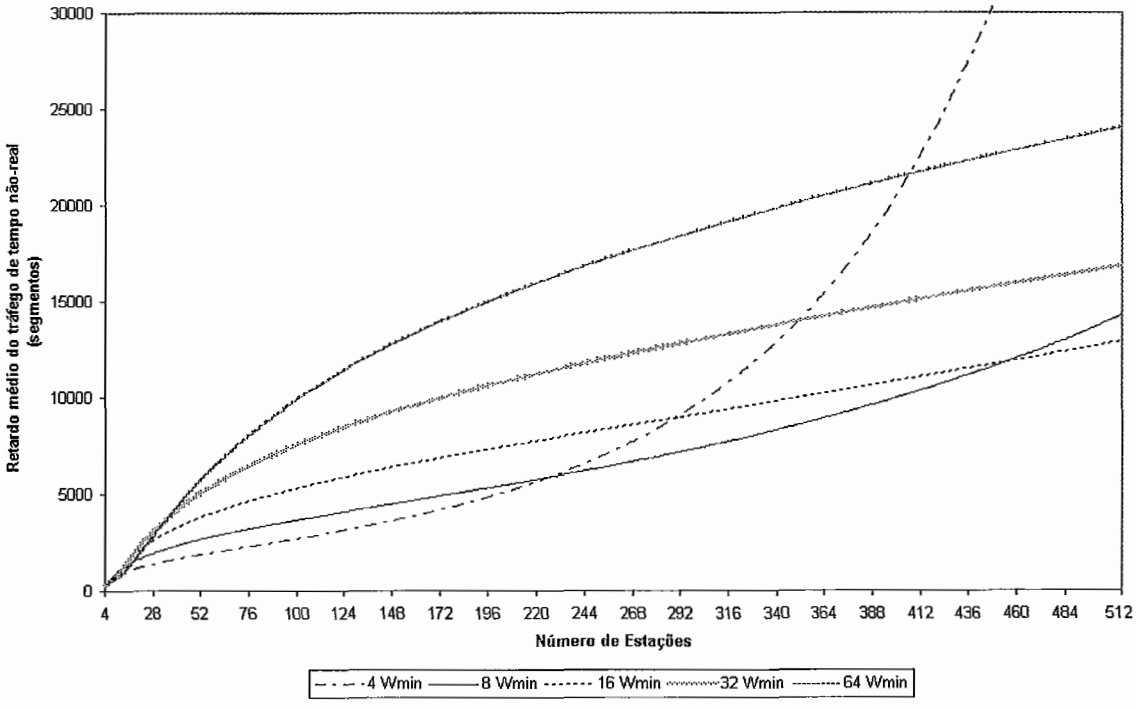


Figura 5.6: Retardo médio total do tráfego de tempo não-real versus o número de estações, que disputam o acesso ao meio, com diferentes tamanhos da janela inicial de *backoff*.

tentativa de transmissão é menor, visto que a janela de *backoff* dobra após ocorrer uma colisão. Porém, para uma quantidade grande de estações, as mensagens de requisição de largura de banda sofrem muitas colisões, demorando muito a convergir para janelas onde suas chances de colisões são minimizadas, não sendo capaz de resolver o conflito. Ainda resta comentar que o retardo ocasionado com o valor da janela inicial de 64 segmentos é devido apenas às colisões experimentadas pelas mensagens de requisição de largura de banda, visto que as janelas de *backoff* inicial e máxima são as mesmas.

Na figura 5.7, o eixo x indica a quantidade de estações que disputam pelo acesso ao canal de *uplink* no período de contenção, enviando mensagens de requisição de largura de banda, e o eixo y indica a utilização do segmento de contenção para diferentes valores da janela inicial de *backoff*, considerando os parâmetros $\{m = 6, R = 10, \mu = 8\}$. Para cada tamanho da janela inicial de *backoff* há

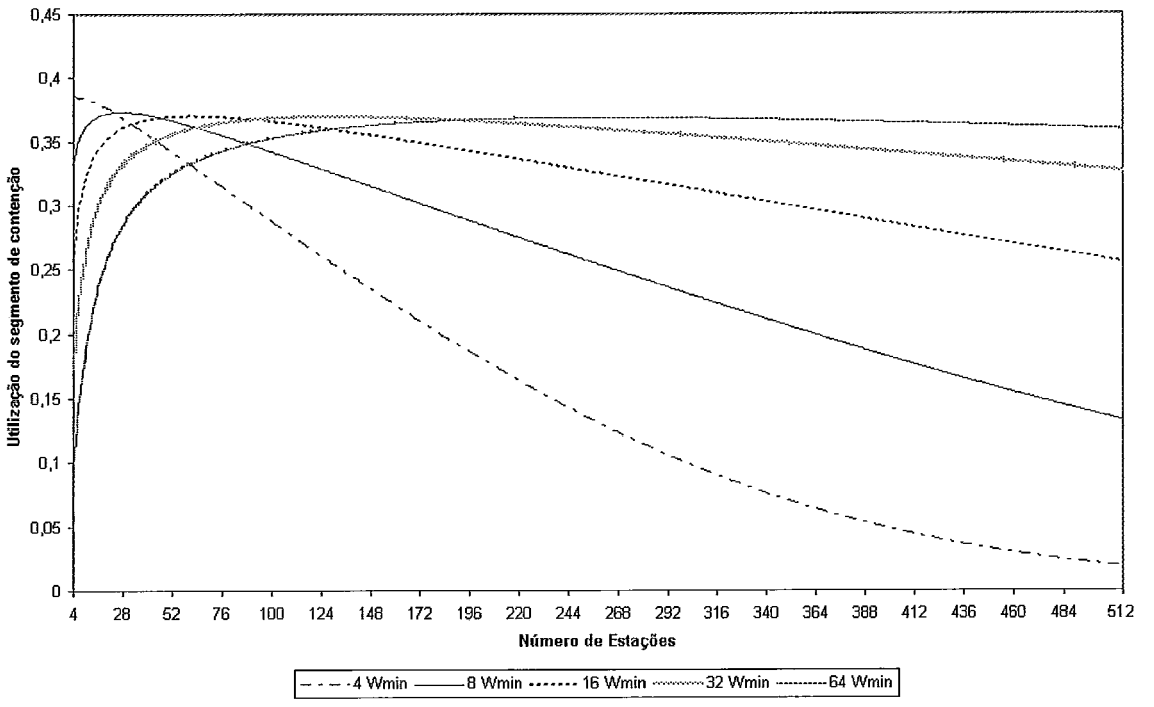


Figura 5.7: Utilização do segmento de contenção versus o número de estações, que disputam o acesso ao meio, com diferentes tamanhos da janela inicial de *backoff*.

uma quantidade de estações que maximiza a utilização do segmento de contenção. Quando o número de estações e o tamanho da janela inicial de *backoff* é pequeno, a utilização do segmento é alta e cai drasticamente com o aumento do número de estações, esse comportamento é devido ao aumento do número de colisões experimentadas pelas mensagens de requisição de largura de banda. Por outro lado, quando o número de estações é pequeno e o tamanho da janela inicial de *backoff* é maior, a utilização do segmento é pequena, ocasionada por um grande número de segmentos vazios, ou seja, onde não ocorre transmissão. A medida que cresce o número de estações a utilização do segmento aumenta, devido a um aumento das mensagens de requisição de largura de banda.

Pelas figuras 5.6 e 5.7, observa-se que os valores que maximizam a utilização do segmento, não fornecem os menores retardos das mensagens, por exemplo, 64 estações com valores de janela inicial de 8 e 32 segmentos, fornecem a mesma utilização do segmento do intervalo de reserva, para diferentes valores do retardo médio

das mensagens. Essa inconsistência ocorre pelo fato de janelas maiores oferecerem maior punição para a transmissão das mensagens, em outras palavras, uma estação que tem o tamanho da janela de *backoff* pequena, e perde a disputa devido a uma colisão, precisará adiar poucos segmentos até a sua próxima tentativa, o que não ocorre com estações que possuem janelas maiores.

Na figura 5.8, o eixo x indica a quantidade de estações que disputam pelo acesso ao canal de *uplink* no período de contenção, enviando mensagens de requisição de largura de banda, e o eixo y indica o retardo médio total das mensagens de tempo não-real para diferentes valores da janela máxima de *backoff*, considerando os parâmetros $\{W_{min} = 8, \mu = 8\, R = 13, 11, 9, 7, 5, 3, 0\}$. Quando o número de estações aumenta de 4 até 512, o retardo também cresce devido ao aumento das colisões das mensagens de requisição de largura de banda. Para pequenos tamanhos da janela máxima de *backoff*, os retardos das mensagens são significativamente menores, entretanto, para a janela de *backoff* máxima igual a 8 com mais de 128 estações, o processo de resolução de contenção não consegue resolver o conflito. A grande diferença dos maiores retardos, oferecidos pelas maiores janelas de *backoff*, em relação as menores deve-se ao fato do adiamento para tentar uma nova oportunidade de transmissão após uma colisão. Visto que a janela de *backoff* dobra o seu tamanho a cada colisão, não limitar o tamanho da janela máxima, implica em aumentos significativos dos retardos das mensagens. Pelo gráfico, observa-se que o valor da janela máxima que minimiza o retardo, para até 320 estações, é dado para 32 segmentos, ou seja, m=5 e R=11.

Na figura 5.9, o eixo x indica a quantidade de estações que disputam pelo acesso ao canal de *uplink* no período de contenção, enviando mensagens de requisição de largura de banda, e o eixo y indica a utilização do segmento de contenção para diferentes valores da janela máxima de *backoff*, considerando os parâmetros $\{W_{min} = 8, \mu = 8\, R = 13, 11, 9, 7, 5, 3, 0\}$. Cada janela máxima encontra sua utilização máxima para um determinado número de estações, entretanto, esse valor não reflete a otimização do retardo médio das mensagens. Mesmo com a baixa utilização da janela máxima de 32 segmentos, observa-se pelo gráfico do retardo, que os valores para essa janela são os que oferecem o melhor retardo médio para o maior "range"de

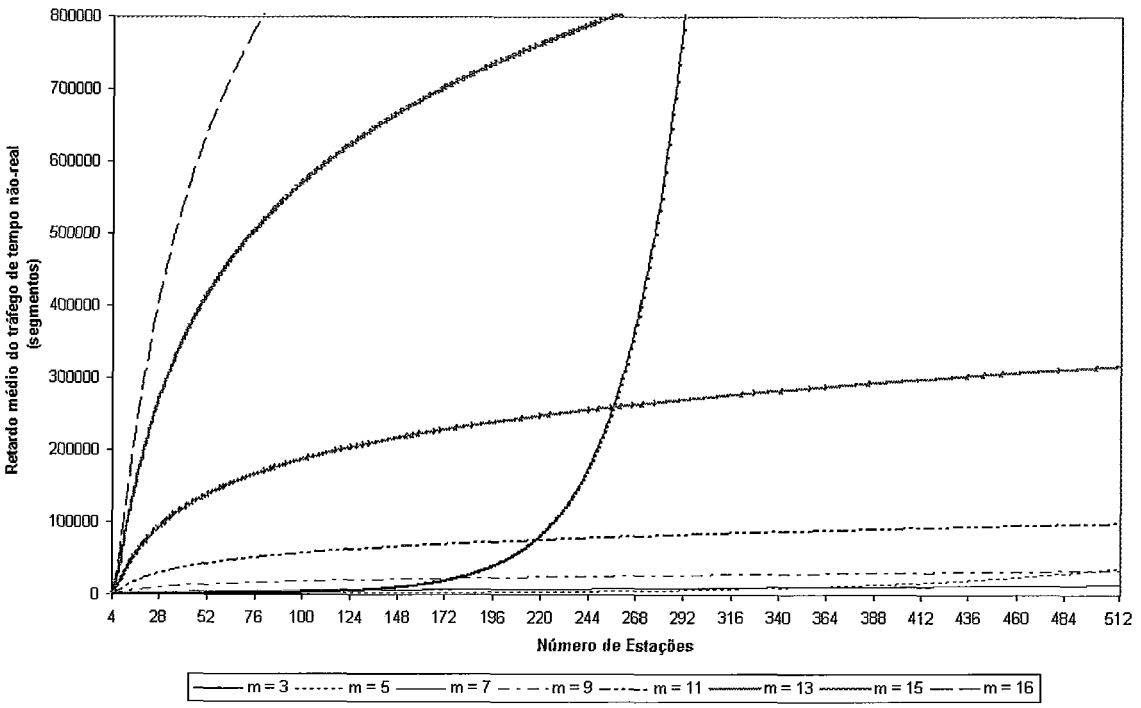


Figura 5.8: Retardo médio total do tráfego de tempo não-real versus o número de estações, que disputam o acesso ao meio, com diferentes tamanhos da janela máxima de *backoff*.

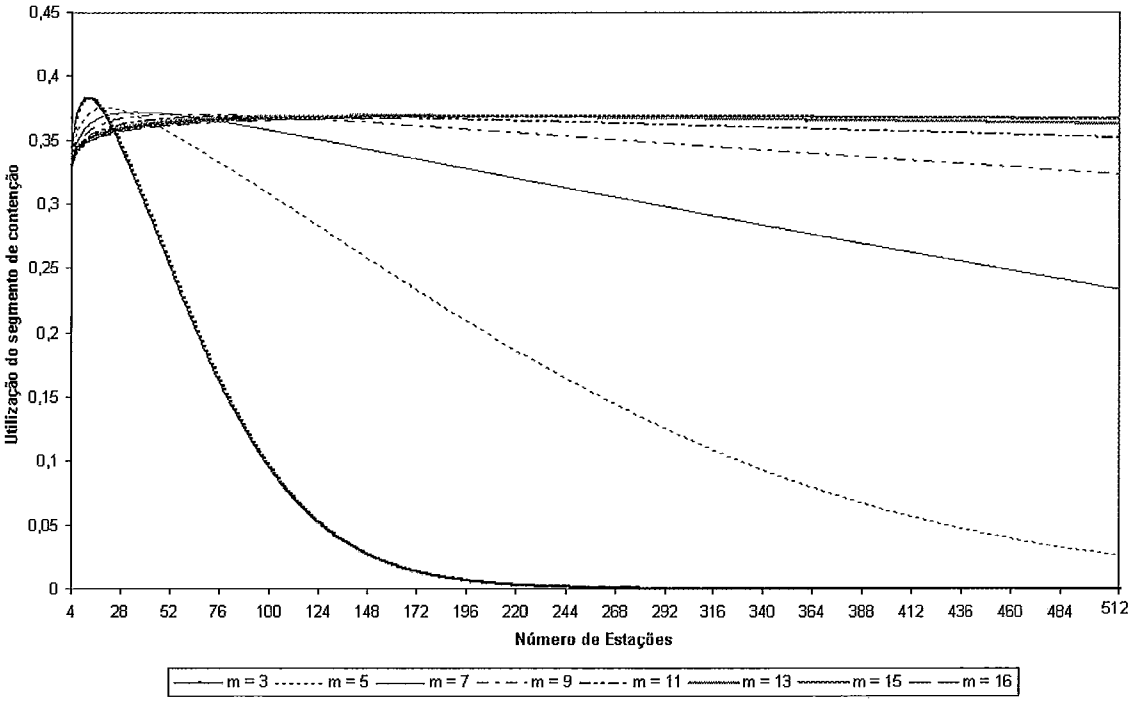


Figura 5.9: Utilização do segmento de contenção versus o número de estações, que disputam o acesso ao meio, com diferentes tamanhos da janela máxima de *backoff*.

número de estações. Esse comportamento é causado pela quantidade de segmentos que uma estação, com janela máxima grande, tem que esperar para transmitir sua mensagem de requisição, no caso de colisões consecutivas.

Na figura 5.10, o eixo x indica a quantidade de estações que disputam pelo acesso ao canal de *uplink* no período de contenção, enviando mensagens de requisição de largura de banda, e o eixo y indica o retardo médio total das mensagens de tempo não-real para diferentes valores do período de contenção, reservado a oportunidades de requisição de largura de banda, considerando os parâmetros $\{m = 6, R = 10\}$. A janela inicial de *backoff* é do mesmo tamanho do período de contenção. Quando o número de estações aumenta, o retardo das mensagens também cresce devido a ocorrência de colisões das mensagens de requisição de largura de banda. O retardo das mensagens é influenciado pelo período de contenção, onde maiores períodos de contenção oferecem menores retardos. Porém, aumentar o período de contenção, implica em diminuir os segmentos destinados à transmissão de dados. Sendo assim,

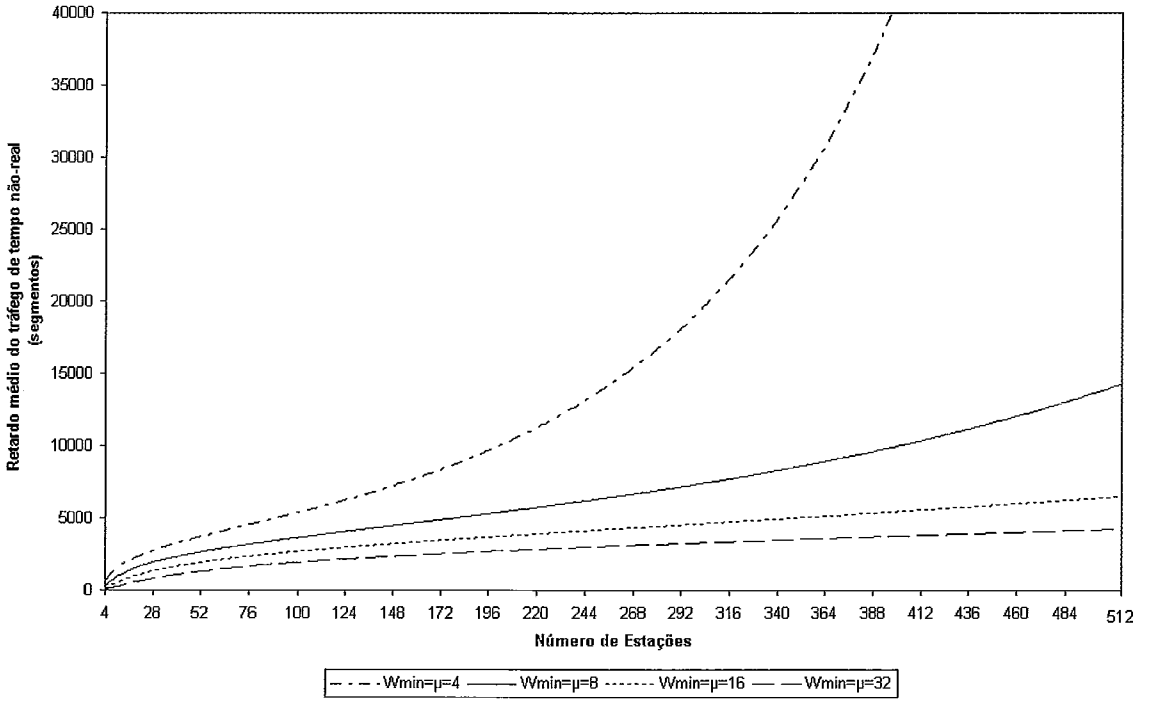


Figura 5.10: Retardo médio total do tráfego de tempo não-real versus o número de estações, que disputam o acesso ao meio, com diferentes tamanhos do período de contenção, reservado para o envio das mensagens de largura de banda.

escolher menores períodos de contenção é desejável, com nível de retardo aceitável.

Na figura 5.11, o eixo x indica a quantidade de estações que disputam pelo acesso ao canal de *uplink* no período de contenção, enviando mensagens de requisição de largura de banda, e o eixo y indica a utilização do segmento de contenção para diferentes valores do período de contenção, reservado às oportunidades de requisição de largura de banda, considerando os parâmetros $\{m = 6, R = 10\}$. Encontra-se a utilização máxima para cada tamanho do período de contenção, contudo, novamente, esse valor não reflete a otimização do retardo médio das mensagens. Para o mesmo valor da utilização com 64 estações e $\mu = 8$ e 32, encontra-se diferentes valores do retardo das mensagens, porém com a introdução de grande *overhead*, ocasionado pelo aumento do período de contenção.

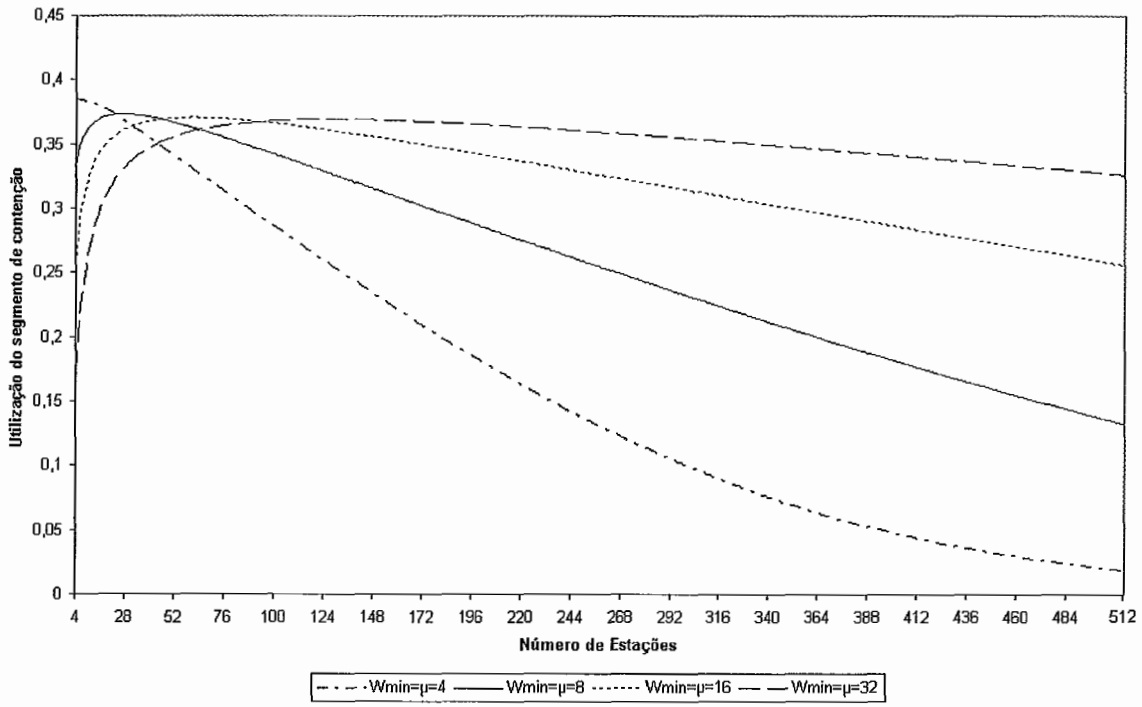


Figura 5.11: Utilização do segmento de contenção versus o número de estações, que disputam o acesso ao meio, com diferentes tamanhos do período de contenção, reservado para o envio das mensagens de largura de banda.

5.3 Validação do Modelo

Para validar o modelo analítico apresentado no capítulo anterior, foi utilizada a ferramenta de modelagem e simulação NS-2 (*Network Simulator*) [35]. Para que fosse possível simular o padrão IEEE 802.16, foi utilizado o módulo para a camada MAC do NS-2 desenvolvido por Borin et. al. [36]. Cada simulação foi rodada 10 vezes com sementes diferentes para gerar o intervalo de confiança de 95% usando o método de replicação. As figuras mostram a média obtida e o intervalo de confiança de 95%.

Os experimentos de simulação têm como objetivo analisar o comportamento da modelagem analítica em uma rede com canal em condições ideais, ou seja, sem perdas ou alterações das mensagens. O cenário de simulação consiste em uma BS com as estações uniformementes distribuídas ao seu redor. Este cenário não tem a intenção de ser representativo para redes operacionais. O objetivo é analisar o mecanismo de acesso ao meio e a alocação de segmentos para diferentes intensidades de tráfego e número de estações. Foi usado fontes CBR para simular o tráfego dos dois tipos de fluxo de serviço. Isto foi necessário, para facilitar a análise dos resultados obtidos através do modelo analítico.

Em todos os cenários simulados supõe-se a presença de um mecanismo de controle de admissão para que os resultados não sejam influenciados por um número excessivo de conexões na rede. Para evitar que o escalonamento nas SSs interfira na avaliação do mecanismo de escalonamento na BS, cada SS possui apenas um fluxo de tráfego. Os parâmetros de configuração da rede são reportados na tabela 5.4.

A modelagem do retardo das mensagens de tempo real é validada usando um cenário de simulação com 1 BS e com o número de SSs variando de 1 até 84. As estações tem fluxo no sentido de *uplink*, com taxa de dados de 64 kbps e mapeadas para o serviço UGS. O intervalo de *grants* é de 10 ms, pois, de acordo com o padrão IEEE 802.16, a BS deve alocar *grants* para esse serviço em intervalos iguais aos intervalos em que a aplicação gera os pacotes. Como observado na tabela 5.4, o quadro de *uplink* consiste de 50 segmentos, onde 8 são destinados a contenção,

Parâmetro	Valores
Largura de banda do canal	40 Mbps
Tempo de quadro	5 ms
Tamanho do segmento	250 bytes
Tempo de segmento	0.05 ms
Janela inicial de *backoff*	8 segmentos
Janela máxima de *backoff*	64 segmentos
Número de tentativas no estágio máximo	10
Número de segmentos de contenção por MAP	8
Tamanho das mensagens	1 segmento
Tamanho do quadro de *uplink*	50 segmentos
Técnica de duplexação	TDD

Tabela 5.4: Parâmetros da simulação.

ou seja, ao envio de mensagens de requisição de largura de banda, e 42 para a transmissão de dados efetiva. Portanto, considerando que cada estação precisa de 1 segmento para transmitir a mensagem gerada a cada 10 ms e os parâmetros deste cenário, cada estação gera uma carga de 0,012 no canal de transmissão.

A figura 5.12 apresenta o retardo médio das mensagens de tempo real, com a carga variando de 0,01 até 1, obtidos por simulação e através da modelagem analítica. O retardo do tráfego UGS não foi afetado com o aumento da carga oferecida, gerada com o aumento do número de estações, o que indica que o modelo pode adequar-se aos requerimentos de retardo do serviço UGS.

O cenário de simulação para o tráfego de tempo não-real consiste de 1 BS, 8 estações com serviço BE, com taxa de dados de 200 kbps, e o número de estações com serviço UGS, com taxa de dados de 64 kbps, varia de 0 até 66. O intervalo de *grants* para o serviço UGS é de 10 ms. Todas as estações tem fluxo no sentido de *uplink*. Considera-se que as estações que geram fluxo de tempo real, precisam de 1 segmento para transmitir a mensagem gerada a cada 10 ms e as estações que geram fluxo de tempo não-real, sempre têm mensagens prontas a enviar em cada quadro e

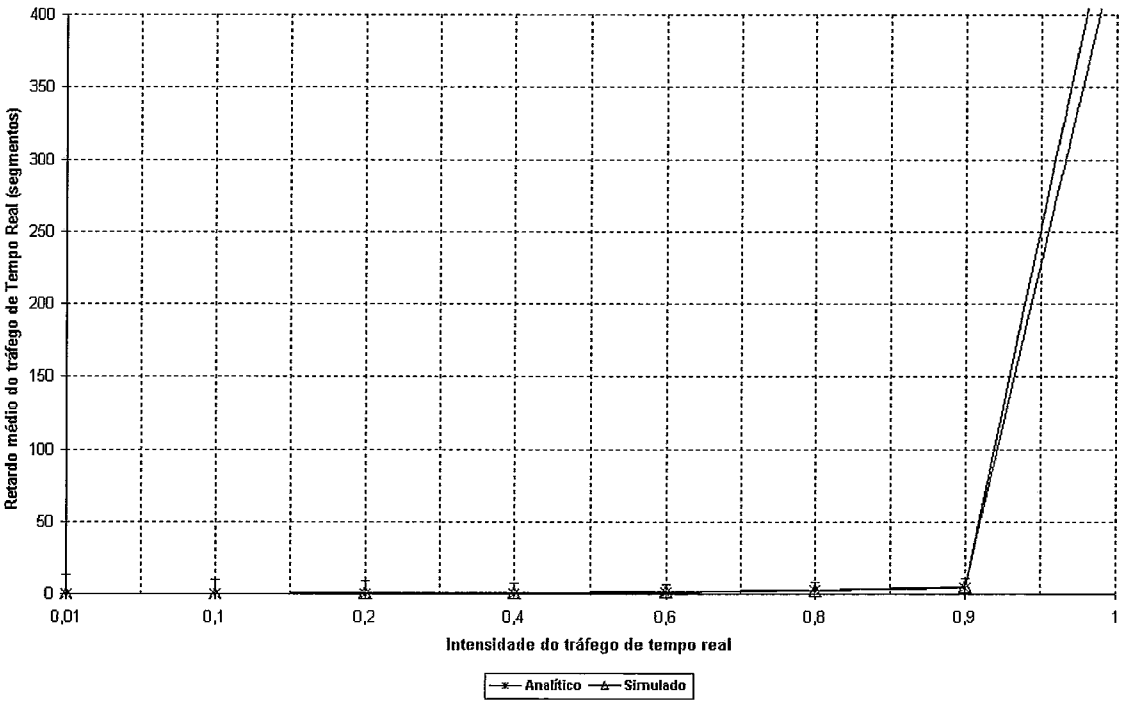


Figura 5.12: Retardo médio do tráfego de tempo real.

também precisam de 1 segmento para transmiti-las. Assim, sob os parâmetros deste cenário, cada estação de tempo real gera uma carga de 0,012 no canal de transmissão, e as de tempo não real geram uma carga de 0,024 no canal de transmissão.

Como função de $\rho_2 = \lambda_2 \nu_2$, o retardo médio do tráfego de tempo real associado com a carga oferecida do tráfego de tempo não-real pode ser avaliado, devido ao fato de que o retardo do tráfego de tempo não-real está relacionado a carga oferecida pelo tráfego de tempo real. Seja $\rho_1 = 0.0, 0.24, 0.36, 0.48, 0.60, 0.71, 0.79$, para $\rho_2 = 0.2$ fixado. O retardo médio do tráfego de tempo não-real é plotado como mostrado na figura 5.13, onde podemos ver que, para a carga do tráfego de tempo não-real oferecida, quanto maior a intensidade do tráfego de tempo real, maior será o retardo do tráfego de tempo não-real. Isso ocorre devido ao fato do tráfego de tempo real ter disciplina de prioridade HOL com interrupção sobre o tráfego de tempo não-real. Portanto, os requisitos de retardo dele são obtidos à custa do retardo do tráfego de tempo não-real.

A modelagem do retardo das mensagens de requisição de largura de banda é

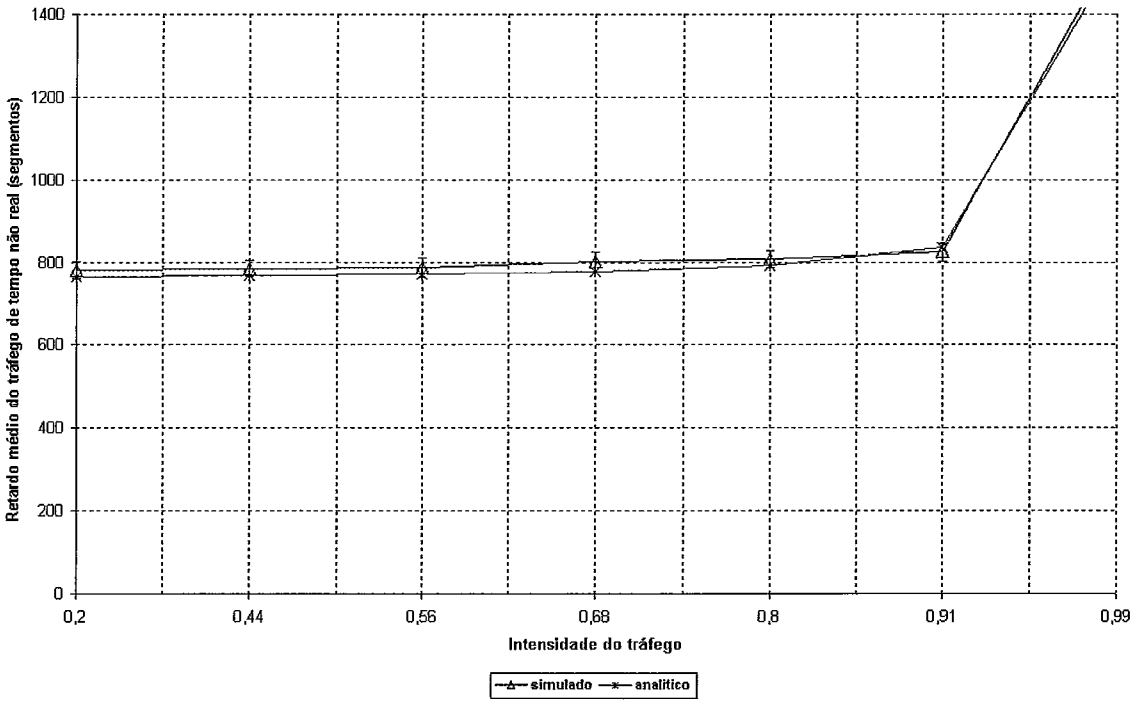


Figura 5.13: Retardo médio do tráfego de tempo não-real versus a carga total oferecida ($\rho = \rho_1 + \rho_2$).

validada usando um cenário de simulação com 1 BS e com o número de SSs variando de 2 até 16. As estações tem fluxo no sentido de *uplink*, com taxa de dados de 200 kbps e mapeadas para o serviço BE.

As figuras 5.14,5.15,5.16,5.17 mostram o retardo médio das mensagens de requisição para quatro diferentes tamanhos da janela inicial de *backoff*. Está claro que o melhor retardo de tempo é obtido para menores valores de W_{min} quando o número de estações é pequeno. Ao contrário, quando o número de estações aumenta, um menor retardo é alcançado com maiores valores de W_{min}. Porém, como visto na seção anterior, para um valor grande de W_{min}, as colisões ocorrem com menor freqüência, visto que elas têm contadores de *backoff* maiores. Ao mesmo tempo, essas estações tem que esperar mais por segmentos de contenção, aumentando assim o retardo das mensagens de requisição de largura de banda. Observa-se ainda, que quando o número de estações aumenta, grandes valores de W_{min} implicam em retardos maiores. Isso se deve ao aumento das colisões e, conseqüentemente, a uma

maior espera para transmissão da requisição devido ao grande valor da janela inicial. Vale ainda ressaltar que, quando ocorre uma colisão o valor de "adiamento"da transmissão é duas vezes o valor da janela inicial e assim sucessivamente.

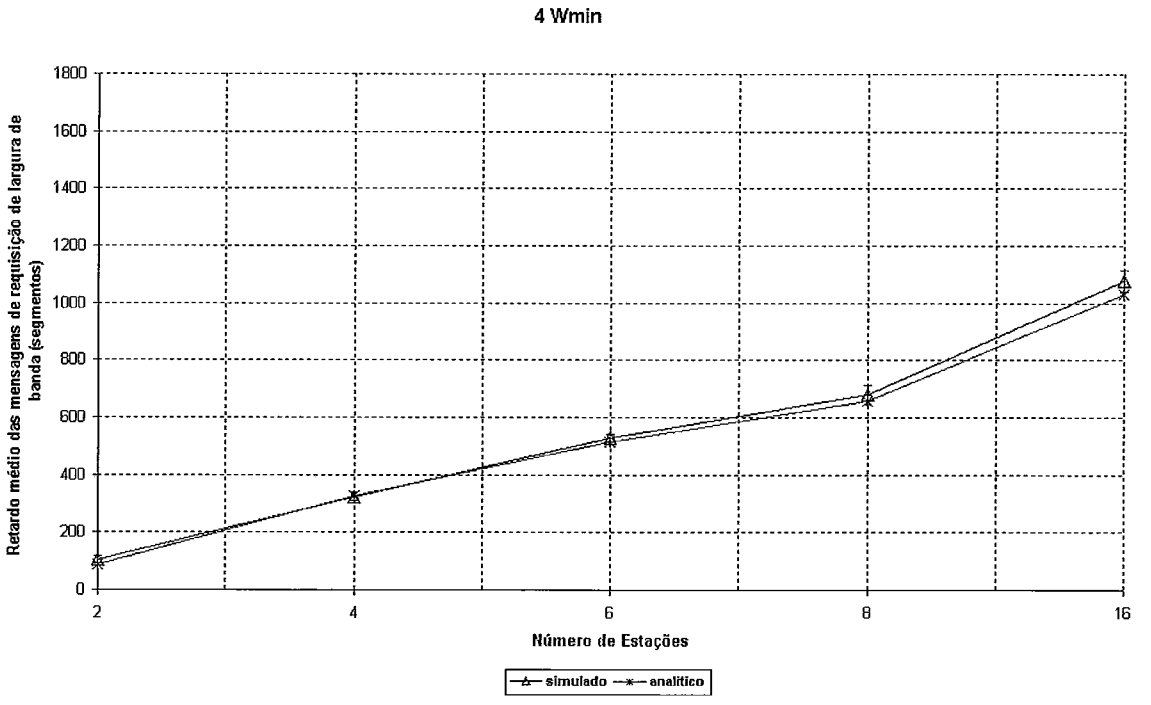


Figura 5.14: Retardo médio do tráfego de tempo não-real para uma janela inicial de *backoff* igual a 4 segmentos.

Como pôde ser observado nessa seção, os resultados obtidos por simulação aproximam-se dos resultados analíticos, o que comprova a viabilidade da utilização do modelo analítico para o que ele se propõe.

5.4 Considerações Finais

Este capítulo apresentou alguns exemplos numéricos do retardo total do sistema para a modelagem analítica proposta. Foi observado o comportamento de diferentes fluxos de prioridades distintas, para a alocação das mensagens de dados, onde a modelagem diferenciou eficientemente essas classes de tráfego. Investigou-se o comportamento do retardo total das mensagens de requisição de largura de banda e a

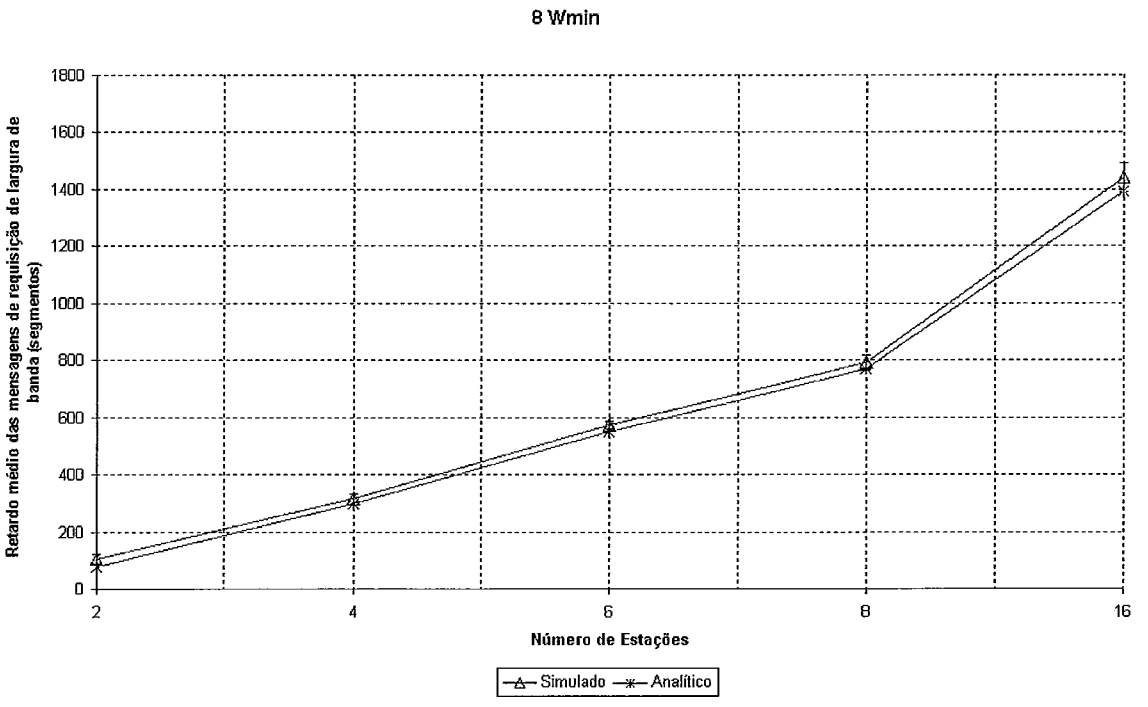


Figura 5.15: Retardo médio do tráfego de tempo não-real para uma janela inicial de *backoff* igual a 8 segmentos.

sua relação com a utilização do segmento de contenção, sob diferentes parâmetros do sistema. Por fim, a modelagem analítica foi validada através de simulação. No próximo capítulo serão apresentadas as conclusões e perspectivas para trabalhos futuros.

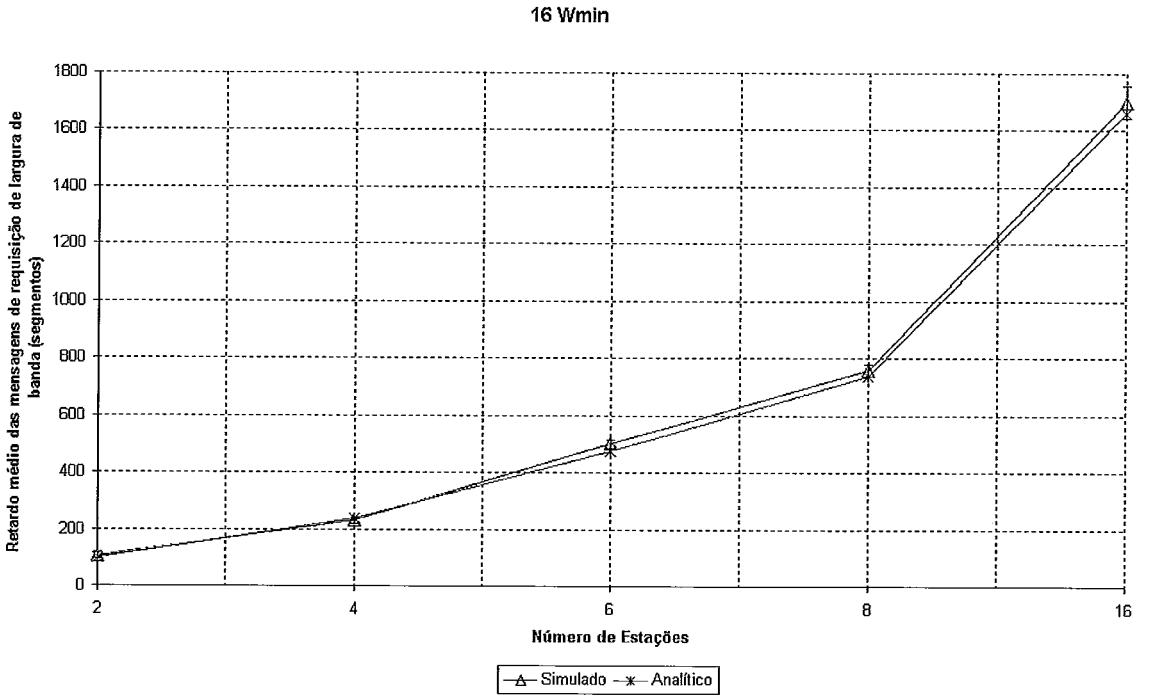


Figura 5.16: Retardo médio do tráfego de tempo não-real para uma janela inicial de *backoff* igual a 16 segmentos.

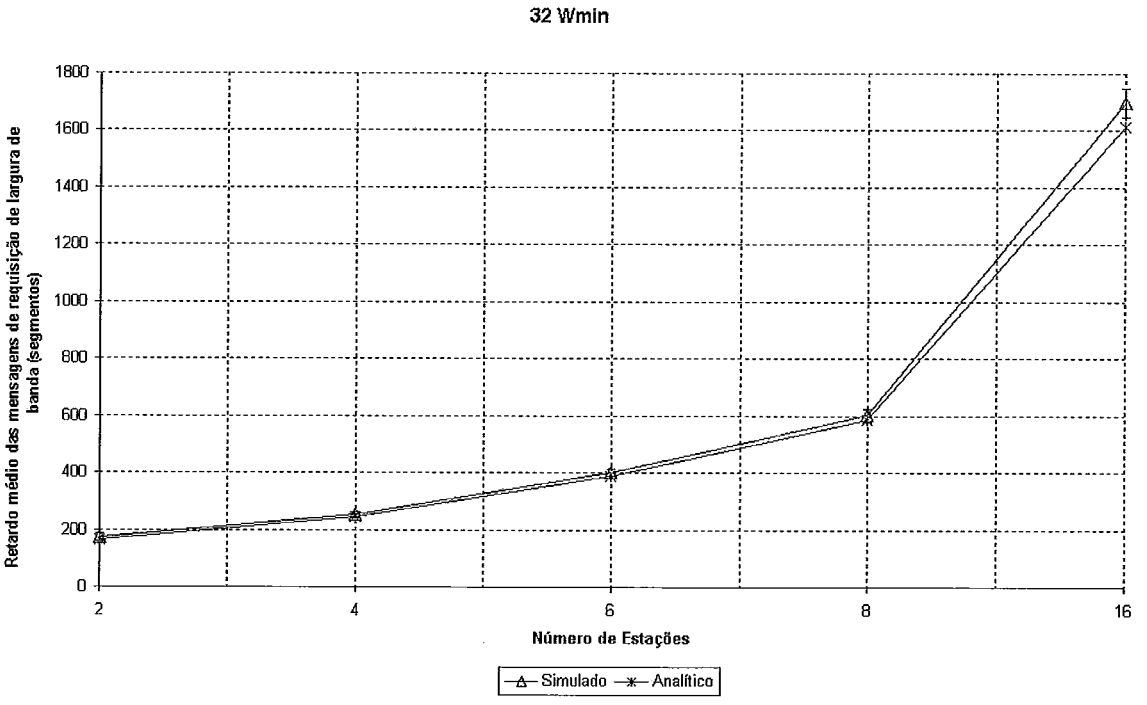


Figura 5.17: Retardo médio do tráfego de tempo não-real para uma janela inicial de *backoff* igual a 32 segmentos.

Capítulo 6

Conclusão e Perspectivas para Trabalhos Futuros

ESTE capítulo conclui o trabalho realizado consolidando os resultados expostos anteriormente e extraindo as conclusões e observações relevantes. Por fim, também são realizadas algumas perspectivas para trabalhos futuros.

6.1 Conclusão

A análise de desempenho através da modelagem analítica constitui uma técnica de fundamental importância dentro do processo de avaliação de desempenho, essa técnica possibilita uma representação e/ou uma análise numérica do problema representado, permitindo assim a realização de uma avaliação de desempenho consistente e rápida. Além disso, quando o sistema a ser representado é muito complexo, essa técnica possibilita uma representação aproximada desse sistema, devido à necessidade de simplificações nessa representação para torná-la numericamente tratável. Com isso, desenvolver um modelo analítico que represente as características do protocolo MAC do padrão IEEE 802.16, não é uma tarefa fácil.

Como foi descrito na seção 3.4, vários trabalhos na literatura investigam o impacto dos mecanismos de escalonamento no desempenho dessas redes, porém esses trabalhos desenvolvem a avaliação de desempenho através de resultados obtidos por simulação. Além disso, alguns trabalhos propõem modelos analíticos para essa avaliação, mas com alguma alteração do protocolo de acesso ao meio do padrão IEEE 802.16.

Dentro desse contexto, foi proposto e apresentado um modelo analítico, utilizando a teoria de filas e cadeia de Markov, para representar o comportamento do protocolo da camada MAC do padrão IEEE 802.16, em termos do retardo fim-a-fim do sistema. O modelo proposto permite a análise de desempenho, das mensagens de requisição de largura de banda e das mensagens de dados, cujos recursos foram alocados através do processo de contenção ou através de garantias pré-estabelecidas. O modelo de requisição de largura de banda é baseado em uma cadeia de Markov, a qual é apropriada para modelar a disputa que ocorre no processo de *backoff* de forma independente. O modelo de alocação dos dados, é baseado na teoria de filas, através de uma fila M/G/1 com prioridade, aplicado nas mensagens com garantias pré-alocadas. Para garantir a prioridade dessas mensagens mais prioritárias, foi incorporado ao modelo um mecanismo de *leaky bucket*, onde considera-se que mensagens virtuais para o tráfego mais prioritário, são sempre servidas antes das demais mensagens. Foram derivadas equações de forma fechada para o retardo total das

mensagens de dados e para as mensagens de requisição de largura de banda.

Os resultados obtidos através da modelagem analítica verificaram as características do protocolo MAC do padrão IEEE 802.16. Pôde-se observar que mesmo para um maior tráfego de mensagens de tempo não-real, o modelo consegue diferenciar eficientemente as classes de tráfego, garantindo menor tempo de espera na fila para as mensagens de maior prioridade, como previsto no padrão. Na avaliação de desempenho das mensagens de tempo não-real, observou-se que a janela de contenção inicial e máxima, bem como o período de contenção, tem forte influência sob o retardo total dessas mensagens, visto que o retardo ocasionado pelas mensagens de requisição é preponderante no retardo total das mensagens de dados. Foi também avaliado o que ocorre no processo de disputa no canal, variando o número de estações na rede, para diferentes parâmetros de *backoff*, onde observa-se que para um maior valor da janela de contenção, a probabilidade da estação transmitir diminui, aumentando conseqüentemente o retardo da mensagem de requisição de largura de banda devido à espera pela oportunidade de transmissão da estação. Além disso, foi mostrado que uma melhor utilização do segmento de contenção não implica em menores retardos das mensagens do tráfego de tempo não-real. Por fim, o modelo analítico foi validado via simulação, o que comprova a viabilidade da utilização do modelo analítico para o que ele se propõe.

6.2 Trabalhos Futuros

Como foi descrito nesse trabalho, a modelagem analítica é uma importante técnica para a análise de desempenho das redes sem fio. Dessa maneira, visando explorar melhor esse tema, com o objetivo de melhor representar as redes IEEE 802.16, o presente trabalho explicita algumas perspectivas de trabalhos futuros:

- avaliação de desempenho do padrão IEEE 802.16 através de outras métricas de desempenho, como: vazão, *jitter*, probabilidade de perda;
- avaliar a modelagem e a análise de desempenho com outros tipos de fontes de

tráfego;

- incorporar ao modelo os demais níveis de prioridade previstos pelo padrão IEEE 802.16;
- Verificar a influência do tamanho das mensagens no retardo total do sistema;
- Propor e modelar diferentes mecanismos alternativos ao *backoff*, como por exemplo, protocolo em árvore para a disputa do acesso ao meio, na reserva de largura de banda das mensagens de menor prioridade.

Bibliografia

[1] IEEE 802.15, *IEEE Standard for Information technology - Telecommunications and information exchange between systems - Local and metropolitan area networks - Specific requirements Part 15.1: Wireless Medium Access Control (MAC) and Physical Layer (PHY) Specifications for Wireless Personal Area Networks (WPANs)*, IEEE Std. 802.15, 2002.

[2] IEEE 802.11, *IEEE Standard for Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specification*, IEEE Std. 802.11, 1999.

[3] ETSI High Performance Radio Local Area Network - HIPERLAN. [Online]. Available: <http://portal.etsi.org/bran/kta/Hiperlan/hiperlan1.asp> Acesso em: 2 de set. 2008

[4] IEEE 802.16, *IEEE Standard for Local and Metropolitan Area Networks - Part 16: Air Interface for Fixed Broadband Wireless Access Systems*, IEEE Std. 802.16, Oct. 2004.

[5] ETSI High Performance Radio Metropolitan Area Network - HIPERMAN. [Online]. Available: <http://portal.etsi.org/> Acesso em: 2 de set. 2008

[6] WiMAX. [Online]. Available: <http://www.wimaxforum.org> Acesso em: 2 de set. 2008

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

[8] M. S. Ross, *Simulation.* Academic Press, 2002.

[9] L. Kleinrock, *Queueing Systems, Vol. 1: Theory.* New York, NY: Addison-Wesley, 1975.

[10] Leonard Kleinrock, "On Resource Sharing in a Distributed Communication Environment," *IEEE Transactions on Communications*, pp. 27–34, Jan. 1979.

[11] P. D. M. Júnior, *Modelagem e Análise de um Protocolo de Acesso Alternativo para o Padrão IEEE 802.16 de Redes Metropolitanas sem Fio.* Dissertação de Mestrado, Programa de Engenharia de Sistemas e Computação - PESC/COPPE/UFRJ, Abril 2005.

[12] H. Peyravi, "Medium-Acceess Control Protocols for Space and Satellite Communications: A Survey and Assessment," *IEEE Communications Magazine*, vol. 37, pp. 62–71, Mar. 1999.

[13] F. A. Tobagi, "Multiaccess Protocols in Packet Communication Systems," *IEEE Transactions on Communications*, Apr. 1980.

[14] J. Kurose, M. Schwartz, and Y. Yemini, "Multiple-Access Protocols and Time-Constrained Communications," *ACM Computing Surveys*, vol. 16, no. 1, Mar. 1984.

[15] L. Kleinrock, "On Some Principles of Nomadic Computing and Multi-Access Communications," *IEEE Communications Magazine*, pp. 46–50, July 2000.

[16] N. Abramson, "The ALOHA System - Another Alternative for Computer Communications," *Fall Joint Computer Conf. AFIPS Conf. Proc.*, vol. 37, pp. 281–285, 1970.

[17] L. Kleinrock, "Packet Switching in Radio Channels: Part I - Carrier Sense Multiple-Access Modes and Their Throughput-Delay Characteristics," *IEEE Transactions on Communications*, Dec. 1975.

[18] F. A. Tobagi and L. Kleinrock, "Packet Switching in Radio Channels: Part III - Polling and (Dynamic) Split-Channel Reservation Multiple Access," *IEEE Transactions on Communications*, vol. 24, no. 8, pp. 832–845, Aug. 1976.

[19] L. F. M. de Moraes and A. N. L. Valverde, "Waiting-Time Analysis of a Reservation Access-Control Scheme with message-based priorities," *IFIP International Conference on Data Communication Systems and Their Performance*, pp. 283–296, June 1987.

[20] J. Kurose and K. Ross, *Computer Networking: A Top-Down Approach Featuring the Internet*, 1st ed. Boston, MA, USA: Addison-Wesley Longman Publishing Co., Inc., 2001.

[21] M. Schwartz, *Broadband Integrated Networks*, 1st ed. Upper Saddle River, NJ: Prentice Hall PTR, 1996.

[22] A. Leon-Garcia and I. Widjaja, *Communication Networks*, 2nd ed. New York, NY: McGraw-Hill, Inc., 2003.

[23] L. F. M. de Moraes e A. B. Kropotoff, "Controle de Congestionamento em Redes ATM," Universidade Federal do Rio de Janeiro - COPPE/PESC, Tech. Rep., 1999.

[24] Intel, "Deploying License-Exempt WiMAX Solutions," *White Paper.*

[25] A. Tanenbaum, *Computer Networks*, 4th ed. Upper Saddle River, NJ: Prentice Hall PTR, 2002.

[26] C. Cicconetti, A. Erta, L. Lenzini, and E. Mingozzi, "Performance Evaluation of the IEEE 802.16 MAC for QoS Support," *IEEE Trans. Mobile Comput.*, vol. 6, pp. 26–38, Jan. 2007.

[27] D. Staehle and R. Pries, "Comparative Study of the IEEE 802.16 Random Access Mechanisms," *Next Generation Mobile Applications, Services and Technologies. NGMAST*, pp. 334–339, Sept. 2007.

[28] B. Bhandari, R. R. Kumar, and S.L.Maskara, "Uplink Performance of the IEEE 802.16 Medium Access Control (MAC) Layer Protocol," *IEEE International conference on Personal Wireless Communications*, pp. 23–25, Jan. 2005.

[29] K. Wongthavarawat and A. Ganz, "IEEE 802.16 Based Last Mile Broadband Wireless Military Networks with Quality of Service Support," *IEEE Military Communications Conference (MILCOM'03)*, vol. 2, pp. 779–784, Oct. 2003.

[30] S.-M. Oh and J.-H. Kim, "The Analysis of the Optimal Contention Period for Broadband Wireless Access Network," *Pervasive Computing and Communications Workshops, 2005. PerCom 2005 Workshops. Third IEEE International Conference on*, pp. 215–219, Mar. 2005.

[31] R. Iyengar, P. Iyer, and B. Sikdar, "Analysis of 802.16 based last mile wireless networks," *Global Telecommunications Conference, 2005. GLOBECOM '05. IEEE*, vol. 5, pp. 1–5, 2005.

[32] D.-H. Cho, J.-H. Song, M.-S. Kim, and K.-J. Han, "Performance Analysis of the IEEE 802.16 Wireless Metropolitan Area Network," *Distributed Frameworks for Multimedia Applications, 2005. DFMA '05. First International Conference on*, pp. 130–137, 2005.

[33] G. Bianchi, "Performance Analysis of the IEEE 802.11 Distributed Coordination Function," *IEEE Journal on Selected Areas in Communications*, vol. 18, no. 3, pp. 535–547, Mar. 2002.

[34] B. D. C. D. I. Choi and D. K. Sung, "Performance Analysis of Priority Leaky Bucket Scheme with Queue-Length-Threshold Schedule Policy," *IEEE Procedure Communications*, no. 145, pp. 395–401, 1998.

[35] "The Network Simulator - ns-2," 2002. [Online]. Available: <http://www.isi.edu/nsnam/ns/>. Acesso em: 2 de set. 2008

[36] J. F. Borin and N. L. S. da Fonseca, "Um Módulo para Simulação de Redes WiMAX no Simulador NS-2," *VII Workshop de Desempenho de Sistemas Computacionais e de Comunicação, Anais do Congresso da Sociedade Brasileira de Computação*, pp. 1–15, 2008.

[37] D. Bertsekas and R. Gallager, *Data Networks*, 2nd ed. Upper Saddle River, NJ: Prentice Hall, 1992.

[38] B. R. Haverkort, *Performance of Computer Communication Systems: a model-based approach.* New York: John Wiley and Sons, 1999.

Apêndice A

Sistema M/G/1

A FILA M/G/1 é um sistema com um único servidor, onde os clientes que chegam na fila da estação formam um processo de Poisson com taxa λ e o tempo de serviço para cada cliente é representado pela variável aleatória X, distribuída de acordo com a função distribuição $B(x)$, isto é, $B(x) = Pr\{X \leq x\}$. X tem média $E[X]$ (primeiro momento) e segundo momento $E[X^2]$.

Define-se $E[N]$ como o número médio de trabalho na fila, $E[N_q]$ como o número médio de usuários na fila, e $E[N_X]$ como o número de usuários no sistema. Aplicando a lei de Little's para o servidor único tem-se: $\rho = E[N_X] = \lambda E[X]$ e assumindo $\rho < 1$ para estabilidade. Pode-se então derivar o número médio de usuários na fila para um sistema de fila M/G/1:

$$N_q = \frac{\lambda^2 E[X^2]}{2(1-\rho)} \tag{A.1}$$

Aplicando a lei de Little´s ($E[W] = E[N_q/\lambda)$, obtém-se o tempo médio de espera na fila:

$$W = \frac{\lambda E[X^2]}{2(1-\rho)} \tag{A.2}$$

Incluindo o tempo de serviço em A.1 e A.2, chega-se às seguintes expressões:

$$E[N] = \lambda E[X] + \frac{\lambda^2 E[X^2]}{2(1-\rho)} \tag{A.3}$$

$$E[R] = E[X] + \frac{\lambda E[X^2]}{2(1-\rho)} \tag{A.4}$$

A equação A.3 é freqüentemente referenciada como fórmula de *Pollaczek-Khinchin* (PK). A prova da fórmula de *Pollaczek-Khinchin* dá-se pelo conceito de *tempo médio de serviço residual* [37].

O sistema M/M/1 é o caso particular do sistema M/G/1 onde os tempos de serviço dos clientes são exponencialmente distribuídos, substituindo na equação (A.2), têm-se:

$$W_{M/M/1} = \frac{\rho E[X]}{(1-\rho)} \tag{A.5}$$

Sistema M/G/1 com Prioridades

Considera-se um modelo de um sistema com servidor único no qual os usuários que chegam ao sistema são classificados em P classes de prioridade, numeradas de 1 até P. Assume-se que a classe 1 possui a maior prioridade e a classe P a menor prioridade. Os usuários da classe k chegam de acordo com um processo de Poisson com taxa λ_k. O tempo médio de serviço para a classe k é $E[X_k] = 1/\mu_k$. O segundo momento do tempo de serviço para a classe k é $E[X_k^2]$. Uma importante característica da estratégia de prioridade é que ela pode ser sem interrupção ou com interrupção.

Sistema sem Interrupção

Nesta estratégia é permitido ao cliente que está em serviço acabar o serviço, caso um cliente de mais alta prioridade chegue na fila naquele instante. O cliente de mais alta prioridade que estiver na fila será servido em seguida, quando o servidor

completar o serviço. Será derivada a relação entre o tempo médio de espera de um usuário da classe k e o tempo médio de espera dos usuários das classes de maior prioridade $1, ..., k-1$.

O tempo de espera do cliente da classe k consiste de três componentes:

- O tempo de serviço residual do cliente em serviço, se algum;
- O tempo de servir todos os clientes das classes mais prioritárias e da mesma classe, ou seja, classe $1, ..., k$, presentes no sistema até a chegada do cliente da classe k;
- O tempo para servir os clientes das classes mais prioritárias, que chegam durante o tempo de espera do cliente da classe k.

Têm-se a seguinte equação para o tempo médio de espera na fila dos clientes de classe k:

$$W_k = T_p + \sum_{r=1}^{k} T_r^{'} + \sum_{r=1}^{k-1} T_r^{''} \tag{A.6}$$

onde T_P é o tempo médio de serviço residual do cliente em serviço, $T_r^{'}$ é o tempo gasto para servir todos os usuários da classe r que estão presentes até a chegada do cliente classe r, e $T_r^{''}$ é o tempo gasto para servir todos os usuários da classe r que chegam durante W_k e que são servidos antes do usuário classe k ser servido. A média dos tempos de cada parcela são dadas por [38]:

$$E[T_P] = \sum_{r=1}^{P} \frac{1}{2} \lambda_r E[X^2] \tag{A.7}$$

$$E[T_r^{'}] = \lambda_r E[W_r]/\mu_r = \rho_r E[W_r] \tag{A.8}$$

$$E[T_r^{''}] = \lambda_r E[W_k]/\mu_r = \rho_r E[W_k] \tag{A.9}$$

Substituindo esses resultados na média da equação A.6, tem-se:

$$E[W_k] = \frac{E[T_P]}{(1-\rho_1-...-\rho_{k-1})(1-\rho_1-...-\rho_k)}, \tag{A.10}$$

onde $\rho_k(=\lambda_k/\mu_k)$ é a utilização do sistema para prioridade k. Assume-se que a utilização total do sistema é menor do que 1, ou seja, $\rho_1+\rho_2+...+\rho_P<1$.

Sistema com Interrupções no Atendimento

Nesta estratégia, o serviço de um cliente será interrompido quando um cliente de mais alta prioridade chegar ao sistema. Neste caso, primeiro o serviço do cliente mais prioritário é servido, depois o cliente será novamente servido, a partir do ponto em que ele parou, quando não existirem clientes mais prioritários no sistema. Dessa forma, a presença de clientes de classes menos prioritária não afetará o atraso médio dos clientes de classe mais prioritária. Portanto, cada classe de prioridade pode ser tratada como se fosse a de menor prioridade no sistema e $E[W_1]$ pode simplesmente ser derivada pela formula de PK para a classe 1:

$$E[W_1] = \frac{\lambda_1 E[X_1^2]}{2(1-\rho_1)} \tag{A.11}$$

Os serviços das classes $k=2,...,P$ esperam o tempo de serviço residual do cliente em serviço, se a sua classe é k, e de todos os clientes de classe mais prioritária. O valor médio desse tempo é dado por:

$$E[T_k] = \sum_{r=1}^{k} \rho_r \frac{E[X_r^2]}{2E[X_r]} \tag{A.12}$$

Então, similar a estratégia sem interrupção, o tempo médio de espera considerando o tempo médio de serviço do cliente $(1/\mu_k)$, o tempo médio de serviço dos clientes de prioridades 1 até k, que se encontravam no sistema e o tempo médio de espera relativo aos clientes de maior prioridade (1 até $k-1$) que chegam enquanto o cliente de classe k está no sistema, pode ser obtido através da seguinte equação:

$$E[T_k] = \frac{(1/\mu_k)(1 - \rho_1 - ... - \rho_{k-1}) + E[T_R]}{(1 - \rho_1 - ... - \rho_{k-1})(1 - \rho_1 - ... - \rho_k)} \tag{A.13}$$

onde, $E[T_R]$ é o tempo médio residual dado pela equação (A.14).

$$E[T_R] = \frac{1}{2} \sum_{i=1}^{k} \lambda_i \overline{X_k^2}. \tag{A.14}$$