

Algoritmos para controle de admissão e escalonamento do tráfego uplink em redes IEEE 802.16*

Juliana Freitag Borin, Nelson L. S. da Fonseca

¹Instituto de Computação – Universidade Estadual de Campinas (UNICAMP)
Caixa Postal 6176 – 13084-971 – Campinas – SP – Brasil

{juliana,nfonseca}@ic.unicamp.br

Abstract. *Although the IEEE 802.16 standard defines five service levels to support real-time and bandwidth demanding applications, admission control and scheduling mechanisms are unspecified. This paper introduces an uplink scheduling algorithm and an admission control policy for IEEE 802.16 networks. The proposed solution is fully standard-compliant and can be easily implemented in the base station. Simulation results show that this solution is able to meet the QoS requirements of multimedia applications.*

Resumo. *O padrão IEEE 802.16 define mecanismos de sinalização entre a estação base e as estações clientes e cinco tipos de serviço para o suporte ao tráfego multimídia. Entretanto, políticas para os mecanismos de controle de admissão e escalonamento, fundamentais para a provisão de Qualidade de Serviço, não são definidas. Este artigo apresenta algoritmos para o controle de admissão e o escalonamento do tráfego uplink compatíveis com as especificações do padrão IEEE 802.16 e que podem ser facilmente implementados na estação base. Resultados de simulação mostram que a solução proposta é capaz de garantir os requisitos de QoS dos diferentes tipos de serviço.*

1. Introdução

Para dar suporte aos requisitos de Qualidade de Serviço (QoS) dos diversos tipos de tráfego, o padrão IEEE 802.16 [IEEE Std. 2004], também conhecido como WiMAX, e sua extensão IEEE 802.16e [IEEE Std. 2005] definem cinco tipos de serviço bem como mecanismos para as conexões *uplink* requisitarem banda. O serviço UGS (*Unsolicited Grant Service*) recebe *grants* de tamanho fixo para o envio de dados periodicamente. O serviço ertPS (*extended real time Polling Service*) utiliza um mecanismo de *grants* semelhante àquele usado pelo serviço UGS, com a diferença de que os *grants* alocados periodicamente podem ser usados para o envio de requisição de banda. O serviço rtPS (*real-time Polling Service*) recebe oportunidades para o envio de requisição de banda periodicamente; este serviço tem garantias de latência máxima e taxa mínima. O serviço nrtPS (*non-real-time Polling Service*), assim como o serviço rtPS, recebe oportunidades periódicas para o envio de requisição de banda, porém com menor frequência, bem como garantias de taxa mínima. O serviço BE (*Best-effort*), compartilha *slots* de contenção com o serviço nrtPS para o envio de requisição de banda e não tem quaisquer garantias de QoS.

*O presente trabalho foi realizado com o apoio do UOL (www.uol.com.br), através do Programa UOL Bolsa Pesquisa, processo número 20071211114200, e do CNPq.

Embora esses serviços forneçam a base para a provisão de QoS, o núcleo desta tarefa reside na alocação de recursos realizada pelo mecanismo de escalonamento. A alocação dos recursos nas redes IEEE 802.16 é realizada por três escalonadores. A estação base possui um escalonador para o tráfego *downlink* (da estação base para as estações cliente) e um escalonador para o tráfego *uplink* (das estações cliente para a estação base). O escalonador *uplink* atribui *grants* para as estações cliente que com o auxílio de um escalonador local decidem quais pacotes serão enviados nas oportunidades de transmissão recebidas.

O padrão IEEE 802.16 não define políticas de escalonamento para que, dessa forma, os fabricantes possam diferenciar seus produtos. Para executar a alocação dos recursos, os escalonadores precisam ter informações sobre os requisitos de QoS e a situação das filas das conexões. O escalonador *downlink* e o escalonador das estações cliente (SSs) têm acesso direto às filas das conexões. O escalonador *uplink* localizado na estação base (BS), por outro lado, depende das requisições de banda enviadas pelas SSs para se manter informado sobre as necessidades de cada conexão. Essas requisições, além de aumentarem a carga na rede, podem sofrer atrasos, causados pelo mecanismo de contenção, resultando na entrega de informações desatualizadas.

Assim sendo, um dos maiores desafios no projeto da camada MAC dos equipamentos para redes WiMAX está na implementação do escalonador *uplink* da BS. Independente da política de escalonamento adotada para o tráfego *uplink*, os seguintes aspectos definidos no padrão devem ser levados em consideração:

1. a distribuição dos recursos deve ser feita com base nas requisições de banda enviadas pelas SSs e nos parâmetros de QoS de cada conexão, sendo que conexões distintas que utilizam um mesmo tipo de serviço podem ter diferentes parâmetros de QoS;
2. a alocação de banda deve permitir não só a transmissão de dados, mas também a transmissão de requisições de banda de acordo com o mecanismo de requisição definido para cada tipo de serviço;
3. todos os parâmetros de QoS definidos pelo padrão devem ser garantidos:
 - *maximum latency*: latência máxima entre o envio de um pacote para a interface de rádio frequência da SS e seu recebimento na interface de rede da BS. Este parâmetro representa um compromisso para os fluxos UGS, ertPS e rtPS que não excedem o parâmetro *minimum reserved traffic rate*.
 - *minimum reserved traffic rate*: requisito de taxa mínima dos serviços ertPS, rtPS e nrtPS.
 - *maximum sustained traffic rate*: taxa de pico dos serviços UGS, ertPS, rtPS e nrtPS. A conexão deve ser policiada para que sua vazão média não ultrapasse esse limite.
 - *maximum traffic burst*: rajada máxima que deve ser acomodada para um fluxo de serviço.
 - *unsolicited grant interval*: intervalo entre *grants* sucessivos alocados para um fluxo UGS ou ertPS.
 - *unsolicited polling interval*: intervalo máximo entre *grants* sucessivos alocados para um fluxo rtPS enviar requisição de banda.
 - *tolerated jitter*: variação tolerada na latência máxima de conexões UGS e ertPS.

Para atender as condições 1 e 3, deve-se limitar o número de conexões simultâneas na rede para que se evite a saturação do enlace sem fio, o que demanda a adoção de um mecanismo de controle de admissão em conjunto com o escalonador *uplink*. Embora o padrão IEEE 802.16 mencione a necessidade de se ter um mecanismo de controle de admissão, políticas não são definidas.

Este artigo introduz algoritmos para o controle de admissão e para o escalonamento do tráfego *uplink* em redes IEEE 802.16 e apresenta uma análise do desempenho destes algoritmos obtida através de experimentos de simulação. Os mecanismos atuam de forma integrada para lidar com as três condições mencionadas anteriormente. O controle de admissão regula a carga imposta à rede, de modo que o escalonador seja capaz de prover os requisitos de QoS das conexões aceitas. Uma nova conexão é aceita quando os recursos disponíveis forem suficientes para atender a taxa mínima requisitada pela nova conexão bem como pelas conexões previamente aceitas. O escalonador, por sua vez, distribui os recursos entre as conexões aceitas de forma que cada uma alcance taxa de transmissão dentro do intervalo definido pelos requisitos de taxa mínima e taxa máxima. As conexões de tempo real têm o requisito de latência máxima garantido para o tráfego que não ultrapassa a taxa mínima, como definido no padrão IEEE 802.16. Os requisitos de tamanho máximo da rajada, de *jitter* e de intervalo entre *grants* para envio de dados e de requisições de banda também são garantidos pelo escalonador.

O algoritmo proposto para o escalonador estende os trabalhos anteriores dos autores. Na versão apresentada em [Borin e Fonseca 2008a], os requisitos de taxa máxima e tamanho máximo da rajada não são policiados. Adicionalmente, o escalonador tenta servir todos os pacotes de tempo real antes do seu prazo máximo de envio (*deadline*), o que exige um esquema de controle de admissão complexo para garantir o requisito de latência máxima, ao mesmo tempo em que mantém alta a utilização da rede. A versão apresentada em [Borin e Fonseca 2008b] inclui o policiamento da taxa máxima, usando a técnica de janelas deslizantes, e do tamanho máximo da rajada, através de testes com o tamanho dos *grants* alocados. No escalonador introduzido neste artigo, o policiamento de ambos os parâmetros é feito por um regulador do tipo duplo balde furado (*dual leaky bucket*), o que simplifica o algoritmo bem como sua configuração, dado que não há necessidade de se investigar o tamanho da janela que oferece o melhor desempenho para os diferentes tipos de tráfego. Tanto o escalonador apresentado em [Borin e Fonseca 2008b], como a versão introduzida neste artigo garantem a latência máxima do tráfego de tempo real que não ultrapassa a taxa mínima requisitada. Dessa maneira, além de se ter um mecanismo totalmente compatível com as especificações do padrão, pode-se projetar um algoritmo de controle de admissão simples, como o esquema apresentado neste artigo, sem a necessidade de monitores e estimadores da latência do tráfego de tempo real e cuja meta consiste em garantir que o escalonador tenha recursos suficientes para prover a taxa mínima das conexões admitidas. Com recursos suficientes para a alocação da taxa mínima das conexões, o escalonador pode dar maior prioridade às conexões de tempo real que não alcançaram a taxa mínima sempre que elas têm pacotes com *deadline* próximo de expirar.

Resultados obtidos via simulação mostram que a solução proposta é capaz de atender os requisitos de QoS das conexões *uplink* em situações onde os cinco tipos de serviço são utilizados para transmitir tráfego multimídia. O restante do artigo está organizado da seguinte forma: a Seção 2 discute trabalhos relacionados. As Seções 3 e 4 apresentam

os algoritmos de controle de admissão e de escalonamento, respectivamente. A Seção 5 descreve a metodologia usada nos experimentos de simulação realizados para analisar o desempenho dos mecanismos. A Seção 6 apresenta os resultados obtidos e a Seção 7 traz as considerações finais.

2. Trabalhos Relacionados

Diversas soluções de escalonamento propostas para redes IEEE 802.16 [Hawa e Petr 2002, Wongthavarawat e Ganz 2003, Chen *et alli* 2005, Tarchi *et alli* 2006] combinam políticas de escalonamento clássicas desenvolvidas para redes cabeadas, tais como *Strict Priority*, *Weighted Fair Queueing* (WFQ) e *Earliest Deadline First* (EDF), resultando em escalonadores complexos. Trabalhos mais recentes [Sayenko *et alli* 2008, Bai *et alli* 2008] bem como a solução proposta neste artigo têm investido em idéias de implementação mais simples. Dado que o escalonador *uplink* da BS é executado a cada *frame* e, em determinadas configurações, pode-se ter até 400 *frames* por segundo, soluções simples são fundamentais. Diferente dos escalonadores propostos por Sayenko *et alli* (2008) e Bai *et alli* (2008), o escalonador proposto neste artigo atende à todas as especificações do padrão IEEE 802.16. O escalonador proposto por Sayenko *et alli* não provê garantias de latência máxima. Na proposta de Bai *et alli*, a provisão dos requisitos de latência, taxa mínima e taxa máxima depende de um valor de prioridade calculado pelas SSs. Embora essa técnica resulte em um escalonador menos complexo na BS, ela limita a interoperabilidade entre equipamentos de diferentes fabricantes, dado que todas as SSs devem ser capazes de calcular os valores de prioridade.

O esquema de controle de admissão proposto neste artigo bem como diversas outras propostas encontradas na literatura [Chen *et alli* 2005, Wang *et alli* 2007a, Wang *et alli* 2007b] usam apenas o requisito de taxa mínima das conexões no processo de decisão. Essa abordagem é válida mesmo para as conexões de tempo real, pois, de acordo com o padrão, a latência máxima precisa ser garantida apenas para o tráfego que não ultrapassa a taxa mínima requisitada pela conexão e essa garantia pode ser implementada pelo escalonador. Algoritmos de controle de admissão que incluem garantias de latência para todo o tráfego de tempo real tendem a ser mais complexos, como, por exemplo, o algoritmo apresentado em [Niyato e Hossain 2007]. Dentre as propostas de controle de admissão que usam apenas o requisito de taxa mínima, o esquema proposto neste artigo diferencia-se pelo fato de considerar a sobrecarga gerada pelo tipo de serviço sendo solicitado (como, por exemplo, a sobrecarga gerada pelo mecanismo de requisição de banda). Chandra e Sahoo (2007) também consideram a sobrecarga dos diferentes tipos de serviço, entretanto, a proposta apresentada por eles não inclui o serviço ertPS. Os autores também não detalham o algoritmo utilizado pela rotina *search*, que é parte importante do mecanismo proposto.

3. Controle de admissão

O mecanismo de controle de admissão limita o número de conexões na rede de modo que os recursos disponíveis para o escalonador sejam suficientes para atender os requisitos de QoS das conexões aceitas. A decisão sobre a admissão de novas conexões leva em consideração o requisito de taxa mínima. Uma vez que esse requisito pode ser garantido para a nova conexão bem como para as conexões previamente admitidas, a conexão é aceita. Uma nova conexão é aceita sempre que a seguinte condição é satisfeita:

$$(C_{reserved} + TR_i^{service} \leq C) \quad (1)$$

$TR_i^{service}$ representa a taxa que deve ser garantida para a nova conexão i associada ao tipo de serviço $service$; $C_{reserved}$ representa a capacidade reservada para as conexões já admitidas na rede e é igual a $\sum_{j=1}^n TR_j^{service}$; e C é a capacidade disponível para o escalonador, ou seja, a quantidade de banda que o escalonador pode alocar para transmissão de dados e requisição de banda através de *grants unicast*.

Embora esta seja uma condição de admissão bastante simples, ela atende às especificações do padrão IEEE 802.16 com a vantagem de ser facilmente implementada na estação base. A condição proposta garante recursos suficientes para o escalonador atender o requisito de taxa mínima das conexões UGS¹, ertPS, rtPS e nrtPS bem como a latência máxima das conexões de tempo real cujo tráfego não ultrapassa essa taxa. Como será mostrado nos resultados apresentados na Seção 6, os recursos não utilizados pelos serviços de maior prioridade, como por exemplo, em situações em que o tráfego gerado por uma conexão é menor do que a taxa mínima requisitada, são destinados para as conexões BE, que não passam pelo processo de admissão.

Adicionalmente, a principal diferença entre o esquema proposto neste artigo e outros esquemas de controle de admissão propostos para o padrão IEEE 802.16 está no cálculo do parâmetro $TR_i^{service}$ utilizado na equação 1. Ao enviar um pedido de admissão, a conexão fornece, entre outros parâmetros, o requisito de taxa mínima ($minTR_i$). Entretanto, se considerarmos a sobrecarga gerada pelo tipo de serviço solicitado (com o envio de requisições de banda, por exemplo), essa taxa não será suficiente para atender as necessidades da conexão. Assim sendo, ao invés de considerar apenas a taxa mínima, o parâmetro $TR_i^{service}$ também inclui uma estimativa da sobrecarga gerada pelo tipo de serviço associado à conexão.

Conexões UGS e ertPS recebem *grants* em intervalos definidos pelo parâmetro *unsolicited grant interval* (ugi). O tamanho do *grant* para conexões UGS é fixo e definido de acordo com o requisito de taxa máxima da conexão (que é igual ao requisito de taxa mínima para esse serviço). Para conexões ertPS, embora o tamanho dos *grants* varie de acordo com a quantidade de banda requisitada, o tamanho mínimo a ser garantido deve basear-se no requisito de taxa mínima. Desse modo, para ambos os tipos de serviço, a BS deve garantir a alocação de *grants* com tamanho igual ao valor fornecido pela seguinte equação:

$$grantSize_i = minTR_i * ugi_i \quad (2)$$

A equação 2 fornece o tamanho dos *grants* em bytes, entretanto, na mensagem UL-MAP² os *grants* são descritos em número de *slots*, obtido a partir do cálculo do menor inteiro maior ou igual a $grantSize_i / slotSize$. Esta normalização pode resultar em

¹No caso do serviço UGS, o padrão estabelece que se deve garantir o requisito de taxa máxima e que o requisito de taxa mínima, quando definido, deve ser igual ao valor da taxa máxima. Portanto, sempre que nos referirmos à taxa mínima do serviço UGS, estamos assumindo que ela é igual à taxa máxima.

²No padrão IEEE 802.16, no início de cada *frame*, a BS envia uma mensagem chamada UL-MAP para informar quais SSs poderão transmitir, quando poderão transmitir e o número de *slots* reservados para cada transmissão.

desperdício de banda quando $grantSize_i$ não é um múltiplo do número de bytes que pode ser transmitido em um *slot* ($slotSize$). A taxa real utilizada por uma conexão UGS (ertPS) pode ser, conseqüentemente, maior do que a taxa máxima (taxa mínima) requisitada. Para contabilizar esta sobrecarga, o valor do parâmetro $TR_i^{service}$ para conexões UGS e ertPS é calculado da seguinte forma:

$$TR_i^{UGS,ertPS} = \frac{\lceil \frac{grantSize_i}{slotSize} \rceil * slotSize}{ugi_i + tj_i} \quad (3)$$

Na equação 3, $\lceil \frac{grantSize_i}{slotSize} \rceil$ fornece o tamanho de um *grant* em número de *slots*. Os *grants* devem ser fornecidos em intervalos de duração ugi com *jitter* máximo tj_i (definido pelo parâmetro *tolerated jitter*). Multiplicando-se o número de *slots* em um *grant* pelo tamanho de um *slot* em número de bytes e dividindo-se este valor pela soma $ugs_i + tj_i$ obtém-se a taxa mínima que deve ser garantida para conexões UGS e ertPS.

Além da taxa mínima, conexões rtPS e nrtPS precisam de *grants* periódicos para o envio de requisição de banda. Assim sendo, a taxa que deve ser garantida para estas conexões é igual a soma do requisito de taxa mínima ($minTR_i$) com a taxa usada pelo mecanismo de requisição de banda, conforme apresentado na equação 4.

$$TR_i^{rtPS,nrtPS} = minTR_i + \frac{upSlots * slotSize}{upi_i} \quad (4)$$

onde $upSlots$ representa o número de *slots* usados pelo mecanismo de requisição de banda e upi_i é o valor fornecido pelo parâmetro *unsolicited polling interval*. Conexões nrtPS não definem o parâmetro upi_i , entretanto, o padrão IEEE 802.16 define que este intervalo deve ser da ordem de 1 segundo e pode ser definido pela BS.

Conexões BE são sempre aceitas dado que este serviço não tem requisitos de QoS.

4. Escalonamento

O algoritmo de escalonamento proposto para o tráfego *uplink* utiliza três filas com prioridades diferentes (Figura 1). A fila de baixa prioridade armazena as requisições de banda enviadas por conexões BE. A fila intermediária armazena as requisições de banda enviadas por conexões rtPS e nrtPS. Essas requisições podem migrar para a fila de alta prioridade para que seus requisitos de QoS sejam garantidos. Além das requisições provenientes da fila intermediária, a fila de alta prioridade armazena os *grants* periódicos para envio de dados e os *grants* periódicos para envio de requisição de banda que devem ser escalonados no *frame* seguinte. A BS executa o escalonador a cada *frame* e a agenda de escalonamento é enviada para as SSs na mensagem UL-MAP.

Grants periódicos são gerados e inseridos na fila de alta prioridade em intervalos pré-definidos. Os intervalos entre *grants* UGS e *grants* ertPS são especificados pelas conexões através do parâmetro *unsolicited grant interval*. O intervalo entre *grants* periódicos para envio de requisição de banda é definido pela BS para conexões nrtPS e pelo parâmetro *unsolicited polling interval* para conexões rtPS. Dessa forma, *grants* periódicos para conexões UGS, ertPS, rtPS e nrtPS são garantidos conforme especificado no padrão. Para garantir o requisito de latência máxima, a BS atribui *deadlines* para cada

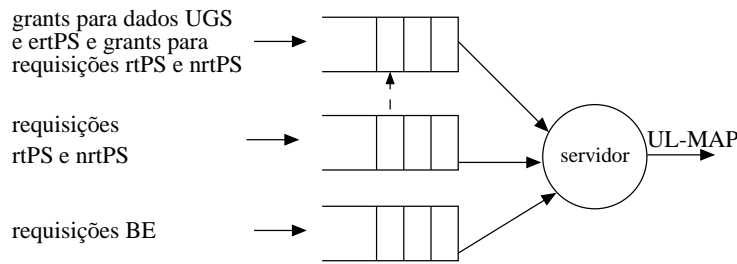


Figura 1. Arquitetura do escalonador proposto

requisição rtPS que entra na fila intermediária. Cada vez que o escalonador é executado, as requisições cujos *deadlines* expiram dois *frames* adiante e cujas conexões ainda não receberam a taxa mínima em uma janela de duração T migram da fila intermediária para a fila de alta prioridade. Dessa maneira, o escalonador garante a latência máxima para o tráfego que não excede a taxa mínima, conforme especificado pelo padrão. Para determinar o *deadline* de uma requisição, é necessário saber o instante de chegada dos pacotes na fila da SS. Dado que a BS não tem acesso à essa informação, considera-se que o pacote que gerou a requisição chegou na fila da SS imediatamente após o envio da última requisição de banda. Assim sendo, o *deadline* de uma requisição é igual ao instante de chegada da última requisição enviada pela conexão correspondente mais o requisito de latência máxima da conexão.

O escalonador garante o requisito de taxa mínima das conexões rtPS e nrtPS em janelas de duração T . Sempre que o escalonador é executado, calcula-se um valor de prioridade para cada requisição de banda na fila intermediária, considerando-se as seguintes informações da conexão: i) requisito de taxa mínima, ii) quantidade de bytes requisitados pela conexão e ainda não atendidos e iii) taxa recebida na janela T atual. Requisições enviadas por conexões que já atingiram a taxa mínima na janela atual recebem valores baixos de prioridade. Para o restante das requisições, quanto menor a taxa atingida pela conexão, maior é a prioridade de suas requisições.

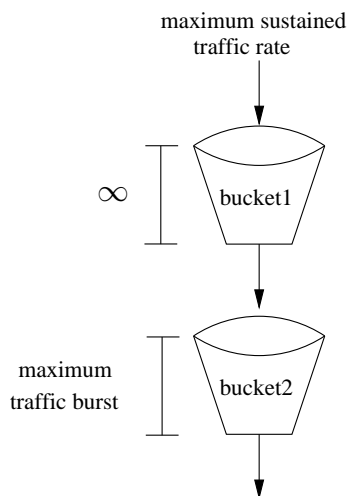


Figura 2. Regulador *dual leaky bucket*

Todos os *slots* alocados para uma SS em um *frame* devem ser combinados em um

ALGORITMO *Scheduling*

1. insere na fila de alta prioridade requisições para os *grants* periódicos que devem ser alocados no próximo *frame*;
 2. CheckDeadline;
 3. CheckMinimumRate;
 4. DistributeFreeResources;
 5. escalona as requisições da fila de alta prioridade começando pela cabeça da fila;
-

PROCEDIMENTO *CheckDeadline*

1. Para cada requisição i na fila intermediária faça
 2. Se $availableBw == 0$ então
 3. break
 4. Se $serviço[CID] == rtPS$ então
 5. $frame = \lfloor (deadline[i] - tempoAtual) \div duraçãoFrame \rfloor$;
 6. Se $frame == 3$ and $TwndTR[CID] < minTR[CID]$
 7. MigrateBWRequest(i)
-

único *grant* na mensagem UL-MAP. Assim sendo, o escalonador garante que o somatório da banda alocada para uma única conexão é menor ou igual ao tamanho máximo da rajada (*maximum traffic burst*). Adicionalmente, o escalonador não aloca banda para uma conexão caso resulte em violação do requisito de taxa máxima (*maximum sustained traffic rate*). O policiamento de ambos os parâmetros é feito por um regulador do tipo duplo balde furado (*dual leaky bucket*), apresentado na Figura 2. O primeiro balde (*bucket1*) tem capacidade infinita e taxa de geração de fichas igual ao parâmetro *maximum sustained traffic rate*. O segundo balde (*bucket2*) tem capacidade igual ao parâmetro *maximum traffic burst* e, a cada *frame*, é preenchido com fichas provenientes do primeiro balde. Dessa forma, antes de alocar um *grant*, o escalonador verifica se o tamanho do *grant* é menor ou igual ao tamanho do *bucket2*.

O Algoritmo *Scheduling* apresenta o mecanismo de escalonamento. Depois de inserir os *grants* periódicos na fila de alta prioridade, o algoritmo verifica quais requisições *rtPS* e *nrtPS* devem migrar da fila intermediária para a fila de alta prioridade (linhas 2 e 3). Na linha 4, o escalonador distribui os *slots* não utilizados pelas conexões de mais alta prioridade entre as conexões BE através da migração de requisições da fila de baixa prioridade para a fila de alta prioridade. No último passo, o escalonador serve todas as requisições armazenadas na fila de alta prioridade.

No procedimento *CheckDeadline*, se há banda disponível, o escalonador usa o procedimento *MigrateBWRequest* para migrar da fila intermediária para a fila de alta prioridade as requisições *rtPS* que satisfazem as seguintes condições: o *deadline* da requisição expira no *frame* que segue o próximo *frame* e a taxa alcançada pela conexão na janela atual ($TwndTR[CID]$, onde CID é o identificador da conexão) é menor do que a taxa mínima requisitada ($minTR[CID]$).

O procedimento *MigrateBWRequest* verifica se a quantidade de bytes solicitados pela requisição sendo migrada ($BR[i]$) está disponível no *subframe uplink*. Caso a quantidade de banda disponível seja menor do que o requisitado, apenas parte da requisição será atendida nesta rodada de escalonamento. O atendimento de parte da requisição também

PROCEDIMENTO *CheckMinimumRate*

1. Para cada conexão i rtPS ou nrtPS faça
 2. $\text{backlogged_tmp}[CID] = \text{backlogged}[CID]$
 3. $\text{TwndTR_tmp}[CID] = \text{TwndTR}[CID]$
 4. $\text{bucket2_tmp}[CID] = \text{bucket2}[CID]$
 5. Para cada requisição i na fila intermediária faça
 6. Se $\text{minTR}[CID] \leq \text{TwndTR_tmp}[CID]$ ou $\text{bucket2_tmp}[CID] = 0$ então
 7. $\text{prioridade}[i] = 0$;
 8. Senão
 9. $\text{prioridade}[i] = \text{backlogged_tmp}[CID] - (\text{TwndTR_tmp}[CID] - \text{minTR}[CID])$;
 10. $\text{TwndTR_tmp}[CID] = \text{TwndTR_tmp}[CID] + \text{BR}[i]$
 11. $\text{bucket2_tmp}[CID] = \text{bucket2_tmp}[CID] + \text{BR}[i]$
 12. $\text{backlogged_tmp}[CID] = \text{backlogged_tmp}[CID] - \text{BR}[i]$
 13. ordene a fila intermediária
 14. Para cada requisição i na fila intermediária faça
 15. Se $\text{availableBw} == 0$ então
 16. break
 17. $\text{MigrateBWRequest}(i)$
-

PROCEDIMENTO *DistributeFreeResources*

1. Para cada requisição i na fila de baixa prioridade faça
 2. Se $\text{availableBw} == 0$ então
 3. break
 4. $\text{MigrateBWRequest}(i)$
-

PROCEDIMENTO *MigrateBWRequest(i)*

1. Se $\text{BR}[i] > \text{availableBw}$ então
 2. $\text{grantSize} = \text{availableBw}$
 3. Senão
 4. $\text{grantSize} = \text{BR}[i]$
 5. Se $\text{grantSize} > \text{bucket2}[CID]$
 6. $\text{grantSize} = \text{bucket2}[CID]$
 7. Se $0 < \text{grantSize} < \text{BR}[i]$ então
 8. crie uma nova requisição j para a conexão CID com $\text{BR}[j] = \text{BR}[i] - \text{grantSize}$
 9. insira a requisição j no final da fila intermediária
 10. $\text{BR}[i] = \text{grantSize}$
 11. mova a requisição i para a fila de alta prioridade
 12. $\text{TwndTR}[CID] = \text{TwndTR}[CID] + \text{grantSize}$
 13. $\text{bucket2}[CID] = \text{bucket2}[CID] - \text{grantSize}$
 14. $\text{backlogged}[CID] = \text{backlogged}[CID] - \text{grantSize}$
 15. $\text{availableBw} = \text{availableBw} - \text{grantSize}$
-

será adiada caso a quantidade de bytes requisitados seja maior do que o tamanho do *bucket2*. Uma vez que a quantidade de banda a ser alocada (*grantSize*) é definida, se *grantSize* é menor do que $BR[i]$, uma nova requisição j é criada com $BR[j]$ igual a $BR[i] - grantSize$ e inserida no final da fila intermediária. A quantidade de banda solicitada pela requisição i é atualizada para *grantSize* e i migra para a fila de alta prioridade. Finalmente, o procedimento atualiza: i) a taxa de tráfego recebida pela conexão na janela T ($TwndTR[*CID*]$), ii) a quantidade de bytes requisitados pela conexão e ainda não atendidos ($backlogged[*CID*]$), iii) o tamanho do *bucket2* da conexão ($bucket2[*CID*]$) e iv) o número de bytes disponíveis no *subframe uplink* (*availableBW*).

O procedimento *CheckMinimumRate* calcula um valor de prioridade para cada requisição na fila intermediária (linhas 8-19). Em seguida, ele ordena a fila intermediária de acordo com os valores de prioridade atribuídos para as requisições (linha 20). No último passo, o escalonador tenta migrar as requisições para a fila de alta prioridade utilizando o procedimento *MigrateBWRequest* e seguindo a ordem de prioridade estabelecida.

O procedimento *DistributeFreeResources* distribui os *slots* não utilizados pelos serviços de mais alta prioridade para as conexões BE migrando as requisições da fila de baixa prioridade para a fila de alta prioridade.

5. Experimentos de simulação

Os mecanismos propostos foram avaliados através de experimentos de simulação utilizando um módulo para simulação de redes WiMAX [Borin e Fonseca 2008c] desenvolvido para a ferramenta ns-2 (*Network Simulator*) [ns 2002].

O cenário de simulação consiste em um conjunto de SSs uniformemente distribuídas ao redor de uma BS. A duração do *frame* é de 5 ms e a capacidade do canal é de 40 Mbps, sendo utilizado o modo de duplexação TDD com razão 1:1 para *downlink* e *uplink*. Cada *slot* é capaz de transportar 100 bytes. Os experimentos utilizam tráfego de voz, voz com supressão de silêncio, vídeo, FTP e WEB que são mapeados para as classes de serviço UGS, ertPS, rtPS, nrtPS e BE, respectivamente.

O tráfego de voz é gerado por uma fonte *on/off* [Brady 1969]. A duração dos períodos *on* e *off* segue uma distribuição exponencial com médias 1,2 e 1,8 segundos, respectivamente. Durante o período *on* são gerados pacotes de 66 bytes a cada 20 ms. O tráfego de voz com supressão de silêncio utiliza a codificação EVRC (*Enhanced Variable Rate Codec*) [3GPP2 1999]. Os pacotes são gerados a cada 20 ms utilizando taxa 1 (171 bits/pacote), taxa 1/2 (80 bits/pacote), taxa 1/4 (40 bits/pacote) ou taxa 1/8 (16 bits/pacote). O tráfego de vídeo é obtido a partir de traces de vídeo reais cujas características são apresentadas na Tabela 1 [Seeling *et alli* 2004]. O tráfego WEB é gerado a partir de uma distribuição híbrida Lognormal/Pareto. O corpo da distribuição, correspondendo à uma área de 0,93, é modelado por uma distribuição Lognormal com média de 27.600 bytes enquanto a cauda é modelada por uma distribuição Pareto com média de 146.300 bytes [Barford e Crovella 1998]. O tráfego FTP é gerado utilizando uma distribuição exponencial com média de 512 KBytes.

O intervalo entre *grants* para os serviços UGS e ertPS é de 20 ms. O intervalo entre *grants unicast* para o envio de requisição de banda é de 20 ms para o serviço rtPS e de 1 s para o serviço nrtPS. O requisito de latência máxima do serviço rtPS é de 100 ms

Tabela 1. Características dos traces de vídeo usados nas simulações

vídeo	taxa de pico (Mbps)	CoV do tamanho dos frames	Hurst
Friends vol4 DVD	1.175520	1.820629	0.818063
Gupta Lecture	2.218080	1.398482	0.777175
Baseball	1.859760	1.284597	0.684708
Lady Tramp DVD	0.910080	1.763723	0.813108
Ice Age DVD	1.058400	1.532578	0.834963
Dreamcatcher DVD	0.932400	1.400108	0.700975
Aladdin	2.242800	1.217128	0.876541

e o requisito de taxa mínima e de taxa máxima varia para cada conexão de acordo com a taxa média do vídeo transmitido. O serviço nrtPS tem requisito de taxa mínima de 200 Kbps e de taxa máxima de 300 Kbps. O serviço BE não tem nenhum requisito de QoS.

O tempo de vida das conexões UGS, ertPS, rtPS e nrtPS é determinado por uma distribuição exponencial com média de 600 s para o tráfego de vídeo e 300 s para os demais tipos de tráfego. Da mesma forma, a taxa de chegada das conexões segue uma distribuição exponencial com médias que variam entre 60 s e 2 s para cada tipo de serviço. Desse modo, na situação de menor carga, chega em média 1 conexão de cada tipo de serviço por minuto, enquanto na situação de maior carga chegam em média 30 conexões de cada tipo por minuto.

O número de conexões BE é fixo em todas as situações simuladas e igual a 20. Dado que as simulações tem duração de 3.600 segundos e as conexões BE não passam pelo processo de admissão, a entrada constante de conexões desse tipo de serviço resultaria em um número exagerado de conexões na rede.

Para cada valor de taxa de chegada de conexões testado, foram rodadas 10 replicações com sementes diferentes para gerar o intervalo de confiança de 95%. Os resultados apresentados na Seção 6 mostram a média obtida e o intervalo de confiança de 95%.

6. Resultados

O primeiro resultado (Figura 3) mostra a influência do aumento da taxa de chegada de conexões na probabilidade de bloqueio dos serviços UGS, ertPS, rtPS e nrtPS. Além disso, ele fornece uma visão geral da situação da rede, em termos do número de conexões aceitas para cada tipo de serviço, sob diferentes condições de carga o que ajuda na análise dos resultados apresentados na sequência. Com exceção da situação com menor carga na rede, onde todos os serviços têm taxa de bloqueio igual a 0, em todas as situações simuladas, o serviço UGS apresenta a menor taxa de bloqueio, seguido pelos serviços ertPS, nrtPS e rtPS. Observa-se, adicionalmente, que quanto maior a taxa de chegada de conexões, maior a distância entre as curvas dos serviços ertPS, nrtPS e rtPS em relação à curva do serviço UGS, que apresenta crescimento mais suave. Esse resultado é reflexo da taxa mínima requisitada pelos diferentes tipos de serviço. Enquanto as conexões UGS têm um requisito de taxa em torno de 26 Kbps, as conexões ertPS requisitam 80 Kbps, as conexões nrtPS têm requisito de taxa mínima de 300 Kbps e as conexões rtPS requisitam entre 200 Kbps e 900 Kbps. Caso o provedor tenha interesse em manter probabilidades

de bloqueio semelhantes para todos os tipos de serviço, é possível estender o mecanismo de admissão para estabelecer graus de utilização para cada tipo de serviço evitando-se, dessa maneira, o favorecimento das conexões que requisitam valores mais baixos de taxa mínima.

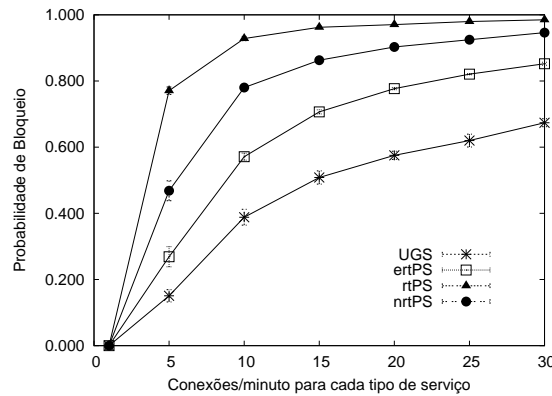


Figura 3. Probabilidade de bloqueio

Os resultados apresentados a seguir comparam os parâmetros de serviço medidos nas simulações com os requisitos de QoS correspondentes definidos para cada fluxo de serviço. O objetivo é verificar a capacidade do controle de admissão e do escalonador para prover QoS para as conexões. Mais especificamente, espera-se que: 1) conexões UGS e ertPS recebam *grants* periódicos conforme o intervalo definido pelo parâmetro *unsolicited grant interval*; 2) o tráfego que não excede a taxa mínima (*minimum reserved traffic rate*) de uma conexão rtPS tenha retardo menor do que a latência máxima requisitada pela conexão; e 3) a vazão média de uma conexão rtPS ou nrtPS não seja menor do que seu requisito de taxa mínima, nem maior do que seu requisito de taxa máxima (*maximum sustained traffic rate*).

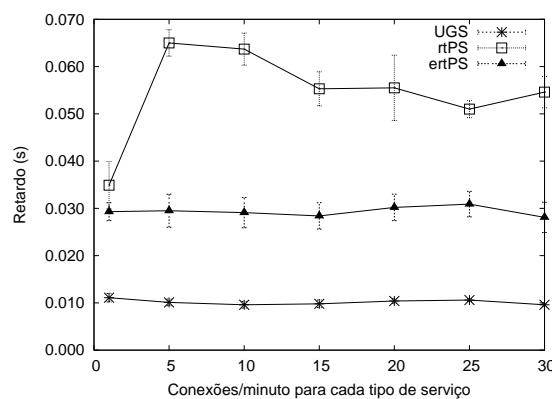


Figura 4. Retardo das conexões UGS, ertPS e rtPS

A Figura 4 apresenta a latência média das conexões UGS, ertPS e rtPS. Para as conexões UGS e ertPS, a latência permanece quase constante em todas as situações, o que indica que as conexões receberam *grants* periodicamente conforme especificado no padrão. A latência das conexões rtPS é menor do que o requisito de 100 ms em todas as situações e apresenta o valor mais alto quando a taxa de chegada é de 5 conexões de cada

tipo de serviço por minuto. Isso se deve ao fato dessa ser a situação com o maior número de conexões rtPS na rede. Embora a taxa de bloqueio seja 0 na situação de mais baixa carga na rede, a taxa de chegada de fontes é de apenas 1 conexão de cada tipo por minuto, o que resulta em um pequeno número de conexões admitidas. Para as situações com taxa de chegada maior do que 5 conexões de cada tipo de serviço por minuto, as conexões rtPS têm menor probabilidade de serem aceitas do que as conexões associadas aos outros tipos de serviço pelas razões mencionadas anteriormente. Situações em que o número de conexões rtPS é elevado apresentam maior taxa de ocupação da fila intermediária, já que as conexões rtPS enviam requisições de banda a cada 20 ms, enquanto as conexões nrtPS enviam requisições de banda a cada 1 s, o que resulta em valores mais altos de latência.

As Figuras 5, 6 e 7 mostram a vazão média das conexões rtPS com requisito de taxa mínima igual a 200Kbps, 500Kbps e 900Kbps, respectivamente. Em todas as situações simuladas, as conexões rtPS obtiveram taxa média dentro do intervalo estabelecido pelos requisitos de taxa mínima e taxa máxima, conforme definido no padrão.

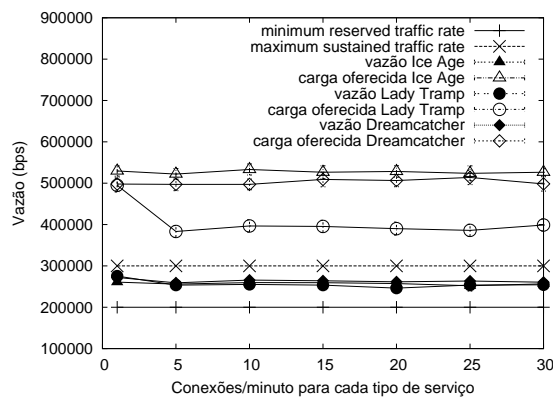


Figura 5. Vazão das conexões rtPS com requisito de taxa mínima de 200 Kbps

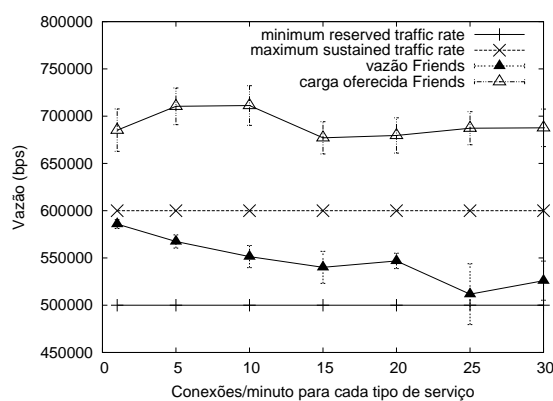


Figura 6. Vazão das conexões rtPS com requisito de taxa mínima de 500 Kbps

A vazão média das conexões nrtPS (Figura 8) se mantém dentro do intervalo definido pelos requisitos de taxa mínima e taxa máxima em todas as situações simuladas. Embora as fontes FTP sejam configuradas para gerar uma carga com média de 300 Kbps, igual ao requisito de taxa máxima do serviço, o mecanismo de controle de congestionamento do protocolo TCP reduz a taxa do tráfego gerado em resposta ao policiamento da

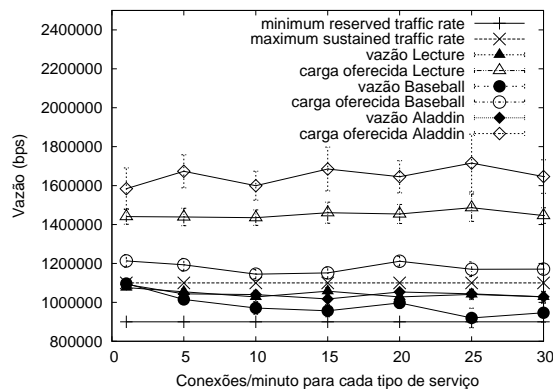


Figura 7. Vazão das conexões rtPS com requisito de taxa mínima de 900 Kbps

taxa máxima feito pelo escalonador. Desse modo, o escalonador consegue servir todo o tráfego gerado pelas fontes em todas as situações de carga simuladas, resultando em valores de vazão maior do que a carga oferecida pelas camadas superiores, devido à sobrecarga imposta pelos cabeçalhos da camada MAC. A carga oferecida foi medida na camada de transporte e reflete a carga enviada para as filas da camada MAC. As estatísticas de vazão foram medidas na camada MAC e mostram a quantidade de tráfego transmitido para a BS, o que permite avaliar o desempenho do escalonador proposto.

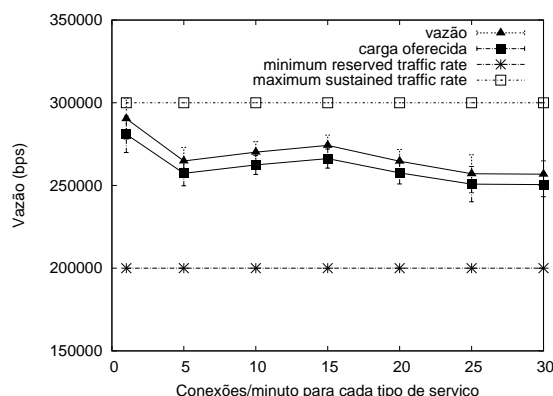


Figura 8. Vazão das conexões nrtPS

A Figura 9 mostra a vazão média das conexões BE para as diferentes taxas de chegada de conexões associadas aos serviços de mais alta prioridade. Assim como acontece com o tráfego nrtPS, o mecanismo de controle de congestionamento do protocolo TCP ajusta a taxa do tráfego gerado de acordo com a banda disponível para o serviço BE, resultando em valores de vazão maiores do que a carga oferecida. Os valores mais baixos de vazão são observados para taxa de chegada de conexões igual a 5 conexões de cada tipo de serviço por minuto. Conforme explicado anteriormente, esta é a situação com maior número de conexões rtPS aceitas. O alto número de conexões rtPS bem como as altas cargas geradas pelo tráfego de vídeo faz com que restem poucos slots para o serviço BE. Por outro lado, quando a taxa de chegada de conexões aumenta e a rede apresenta um maior número de conexões UGS e ertPS, as chances de sobram slots nos frames é maior dado o comportamento on-off do tráfego e, portanto, o serviço BE tem maior acesso aos recursos.

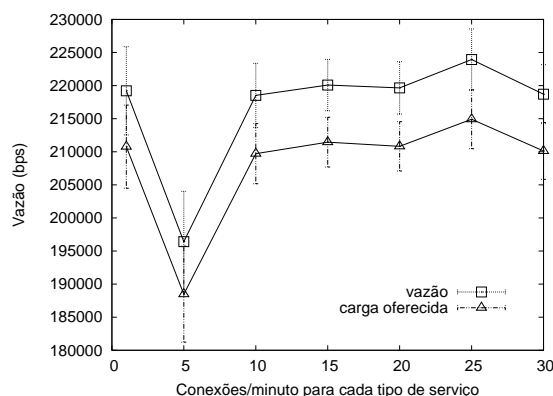


Figura 9. Vazão das conexões BE

7. Conclusões

Este artigo apresentou algoritmos para o escalonamento e o controle de admissão do tráfego *uplink* em redes IEEE 802.16. A admissão de novas conexões leva em consideração o requisito de taxa mínima bem como a sobrecarga gerada pelo tipo de serviço solicitado. O algoritmo de escalonamento proposto fornece garantias de latência e taxa mínima para as conexões aceitas, além de garantir que a taxa máxima e o tamanho máximo da rajada sejam respeitados. Diferente de outras soluções disponíveis na literatura, a solução apresentada neste artigo é compatível com as especificações do padrão IEEE 802.16.

Resultados obtidos a partir de experimentos de simulação mostram que a solução proposta é capaz de prover Qualidade de Serviço para os diferentes tipos de serviço definidos pelo padrão IEEE 802.16. Adicionalmente, o escalonador faz uso eficiente dos recursos disponíveis, alocando para o tráfego de melhor esforço os *slots* não usados pelos serviços de mais alta prioridade.

Referências

- IEEE Standard for Local and Metropolitan Area Networks - Part 16: Air Interface for Fixed Broadband Wireless Access Systems. IEEE Std., Rev. IEEE Std802.16-2004, 2004.
- IEEE Standard for Local and Metropolitan Area Networks - Part 16: Air Interface for Fixed Broadband Wireless Access Systems. Amendment for Physical and Medium Access Control Layers for Combined Fixed and Mobile Operation in Licensed Bands, 2005.
- Barford, P. e Crovella, M. (1998). *Generating representative web workloads for network and server performance evaluation*. Anais do ACM SIGMETRICS Conference, pg. 151-160.
- Bai, X., Shami, A. , e Ye, Y. (2008). *Robust QoS control for single carrier PMP mode IEEE 802.16 systems*. IEEE Transactions on Mobile Computing, vol. 7, no. 4, pg. 416-429.
- Borin, J. F. Borin, e Fonseca, N. L. S. (2008a). *Scheduler for IEEE 802.16 networks*. IEEE Communications Letters, vol. 12, no. 4, pg. 274-276, 2008.

- Borin, J. F. Borin, e Fonseca, N. L. S. (2008b). *A Standard-Compliant Scheduler for WiMAX Networks*. Anais do 4th IEEE Broadband Wireless Access Workshop, pg. 1-5, 2008.
- Borin, J. F., e Fonseca, N. L. S. (2008c). *Simulator for WiMAX networks*. Simulation Modeling Practice and Theory, vol. 16, no. 7, pg. 817-833.
- Brady, P. (1969). *A model for generating On-Off speech patterns in Two-Way conversations*. Bell System Technical Journal, vol. 48, pg. 2445-2472.
- Chandra, S. e Sahoo, A. (2007). *An efficient call admission control for IEEE 802.16 networks*. In Proceedings of the 15th IEEE Workshop on Local & Metropolitan Area Networks, pg. 188-193.
- Chen, J., Jiao, W., Wang, H. (2005). *A Service Flow Management Strategy for IEEE 802.16 Broadband Wireless Access Systems in TDD Mode*. Anais do IEEE International Conference on Communications, pg. 3422-3426.
- 3rd Generation Partnership Project 2 3GPP2 (1999). *Enhanced Variable Rate Codec (EVRC)*. 3GPP2 C. S0014-0, Versão 1.0.
- Hawa, M. e Petr, D. W. (2002). *Quality of Service Scheduling in Cable and Broadband Wireless Access Systems*. Anais do 10th IEEE IWQoS, pg. 247-255.
- Niyato, D. e Hossain, E. (2007). *QoS-aware bandwidth allocation and admission control in IEEE 802.16 broadband wireless access networks: A non-cooperative game theoretic approach*. Computer Networks, vol. 55, no. 11, pg. 3305-3321.
- Sayenko, A., Alanen, O., e Hämäläinen, T. (2008). *Scheduling solution for the IEEE 802.16 base station*. Computer Networks, vol. 52, no. 1, pg. 96-115.
- Seeling, P., Reisslein, M. e Kulapala, B. (2004). *Network Performance Evaluation Using Frame Size and Quality Traces of Single-Layer and Two-Layer Video: A Tutorial*. IEEE Communications Surveys and Tutorials, vol. 6, no. 2, pg. 58-78.
- Tarchi, D., Fantacci, R., e Bardazzi, M. (2006). *Quality of Service management in IEEE 802.16 wireless metropolitan area networks*. Anais do IEEE International Conference on Communications, pg. 1789-1794.
- The Network Simulator - ns-2*. <http://www.isi.edu/nsnam/ns/>, 2002. Acessado em 10 de fevereiro de 2009.
- Wang, L., Liu, F., Ju, Y., e Ruangchaijatupon, N. (2007a). *Admission control for non-preprovisioned service flow in wireless metropolitan area networks*. Anais do Fourth European Conference on Universal Multiservice Networks, pg. 243-249.
- Wang, H., He, B., e Agrawal, D. P. (2007b). *Above packet layer level admission control and bandwidth allocation for IEEE 802.16 wireless MAN*. Simulation Modeling Practice and Theory, vol. 15, no. 14, pg. 266-382.
- Wongthavarawat, K. e Ganz, A. (2003). *IEEE 802.16 based last mile broadband wireless military networks with Quality of Service support*. Anais do IEEE MILCOM'03, pg. 779-784.