

# Uplink Scheduling Algorithms for QoS Support in Broadband Wireless Access Networks

Mikael Gidlund and Gang Wang

**Abstract**—In order to support different types of multimedia applications, the IEEE 802.16 standard defines different service classes with their associated Quality of Service (QoS) parameters. The scheduling algorithm is the crucial point in QoS provisioning over such broadband wireless access (BWA) network and it is important that the scheduling algorithm have a multi-dimensional objective of satisfying QoS requirements of the users, maximizing system utilization and ensuring fairness among users. In this article we present two different scheduling algorithms for the uplink (UL) connection. One is a combination of legacy scheduling algorithms earliest deadline first (EDF) and weighted fair queuing (WFQ). The other proposed algorithm is a cross-layer algorithm that pledges fairness among admitted connections and it also supports all service classes. The proposed scheduling algorithms are compared to several other scheduling algorithms for UL traffic under different mixes of traffic and for various characteristics of the IEEE 802.16 MAC layer such as UL burst preamble, frame length and bandwidth request mechanisms. Simulation results indicate that legacy scheduling algorithms are not suitable for the multi-class traffic in IEEE 802.16 since they do not explicitly incorporate the QoS parameters for the given standard.

**Index Terms**—IEEE 802.16, MAC, scheduling, QoS, fairness, delay.

## I. INTRODUCTION

The increasing interest in wireless broadband communications is a consequence of both rapid growth and the rising importance of wireless communications and multimedia services to end users. In rural areas, broadband wireless access (BWA) represents an economically viable solution to provide last mile access to the Internet, thanks to the easy deployment and low cost of its "light" architecture. Standard activities for BWA are being developed within IEEE project 802, Working Group 16, often referred to as 802.16. The IEEE 802.16 standard is also known in the trade press as Worldwide Interoperability for Microwave Access (WiMAX) [1].

The physical (PHY) layer employs orthogonal frequency division multiplexing access (OFDMA) and supports both fixed and adaptive modulation techniques in the uplink (UL) and in the downlink (DL) directions [2]. Maximum attainable data rates depend upon the modulation schemes used and the condition of the channel. The IEEE 802.16 protocol stack, the medium access control (MAC) layer supports two modes: Point-to-Multipoint (PMP) and Mesh (optional). In the PMP mode, the nodes

are organized into a cellular-like structure, where the base station (BS) serves a number of subscriber stations (SSs) within the same antenna sector in a broadcast manner, with all the SSs receiving the same transmission from the BS. In the mesh mode, the nodes are organized in an ad-hoc fashion and scheduling is distributed among them.

The IEEE 802.16 is designed to support multimedia service via quality of service (QoS) of different service types. Each traffic flow requires different treatment from the network in terms of allocated bandwidth, maximum delay, jitter and packet loss [3]- [5]. *Traffic differentiation* is thus a crucial feature to provide network-level QoS. The standard leaves QoS support features specified for WiMAX networks (e.g., traffic policing and shaping, connection admission control and packet scheduling) open to vendor algorithm design and implementation. One of the most critical issues is the design of a very efficient *scheduling algorithm* which coordinate all other QoS-related functional entities. In the DL, the scheduler has complete knowledge of the queue status, and, thus, may use some classical scheduling algorithms, such as Weighted Round Robin (WRR), Weighted Fair Queuing (WFQ) etc. Priority oriented fairness features are also important in providing differentiated services in IEEE 802.16 networks. Through priority, different traffic flows can be treated almost as isolated while sharing the same radio resource. However, the BS scheduler is non-work conserving, since the output link can be idle even if there are packets waiting in some queues. Indeed, after downlink flows are served in their devoted subframe, no additional downlink flows can be served till the end of the subsequent uplink subframe. Scheduling uplink flows is more complex since the input queues are located in the SSs and are hence separated from the BS. The UL connections work on a request/grant basis. Using bandwidth requests, the uplink packet scheduling may retrieve the status of the queues and the bandwidth parameters.

**Related Work:** In [4], the authors presents an approach based on a fully centralized scheduling (GPC-like) scheme, where a global QoS agent collects all the necessary information on traffic flows, and takes decisions on traffic admission, scheduling, and resource allocation. Based on the complete global knowledge of the system, the deterministic QoS levels can be guaranteed. In terms of the scheduling discipline used for the various classes, both EDF and WFQ are used. Still, the strict priority discipline allow to redistribute bandwidth among its active connections to lowest priority. EDF scheduling has been known for a long time [6] and holds a variety of optimiza-

Gang Wang is with Philips Research in Eindhoven, the Netherlands (email:gang.wang.xu@gmail.com). Mikael Gidlund is with Nera Networks AS in Bergen, Norway. This work was supported in part by Research Council of Norway (NFR).

tions under different scheduling contexts. In [7], Ferrari and Verma proposed EDF as a link scheduler in order to provide delay bounds for real-time communications. In [8], Ruangchaijatupon *et al.* evaluated the performance of the EDF scheduling algorithm for BWA networks. WFQ algorithm is a packet-based approximation of the Generalized Processor Sharing (GPS) algorithm which is an idealized algorithm that assumes that a packet can be divided into bits and each bit can be scheduled separately. WFQ has the nice property of traffic protection, while EDF is known to be optimal in providing delay bounds at a single node [22] and to outperform WFQ in the end-to-end case if per-node traffic shaping is exercised [23]. In [17], Katevenis *et al.* proposed the weighted round robin (WRR) algorithm and it was originally proposed for ATM traffic. In [16], Cicconetti *et al.* implemented the WRR algorithm in IEEE 802.16 MAC layer to evaluate its QoS performance of multi-class traffic. Most of the studied literature about scheduling for IEEE 802.16 have been on the downlink and in uplink scheduling, WFQ and EDF would require computation of virtual start time and finish time at the BS for each packet arriving at the SS.

Most existing schedulers for IEEE 802.16 networks have been designed for *Real Time Polling Service* (rtPS) and *Non Real Time Polling Service* (nrtPS) service rather than for *Best Effort* (BE) services. In [12], Niyato and Hossain proposed an adaptive queue aware uplink bandwidth allocation scheme for rtPS and nrtPS services. The bandwidth allocation is adjusted dynamically according to the variations in traffic load and/or the channel quality. An early work [4] proposed a cross-layer packet scheduling algorithm to provide traffic flows with different classes of service with QoS for time division diversity (TDD) system. In [24], Chen *et al.* devised a scheduling algorithm to jointly serve both UL and DL traffic flows to exploit the dynamic variation of the UL/DL ratio in a TDD system. In [9], Liu *et al.* proposed an algorithm which utilizes users' diversity, however, it does not provide for fairness among users since it allocates slots for one connection in the frame after satisfying the UGS connections requirements. The connections are selected based on a priority function that depicts the QoS requirements of the connections and prioritizes the different service classes through using constant weights. In [12], Niyato and Hossain presented a queuing theoretic (QT) scheduling algorithm which utilizes user diversity for bandwidth allocation. However, the proposed algorithm is a heuristic algorithm for bandwidth allocation instead of an optimal bandwidth allocation algorithm since the complexity of their proposed algorithm may be prohibitive from an implementation point of view.

**Main Contributions:** The overall purpose with this paper is to make a comprehensive performance evaluation of different scheduling algorithms for the uplink in IEEE 802.16 networks. We also present two new scheduling algorithms which will improve the QoS support for IEEE 802.16 network. Our detailed contributions can be summarized as follows:

- We propose one hybrid scheduling algorithm which

has rather low complexity and can be seen as a combination of EDF and WFQ scheduling algorithms. We also propose a cross-layer algorithm which supports all service classes and also integrating a proportional fairness scheme to SS channel quality information to allocate time-slots among connections of the same class.

- We evaluate different scheduling algorithms for the uplink scenario in IEEE 802.16 networks by means of computer simulations and identify performance metrics that will helpfully evaluate the scheduling performance. We also identify open issues and provide suggestions to improve the performance of the evaluated scheduling algorithms.

**Organization of the paper:** The remainder of the paper is organized as follows. Section II briefly introduce the main characteristics of the IEEE 802.16 PHY and MAC layer. In Section III we describe the evaluated scheduling algorithms in more detail, and Section IV describes the simulation set-up and the obtained results. Finally in Section V we conclude our work.

## II. IEEE 802.16 BROADBAND WIRELESS NETWORKS

The basic IEEE 802.16 architecture consists of one base station (BS) and one (or more) subscriber stations (SSs). Both BS and SS are stationary while clients connected to the SS can be mobile. BS acts as a central entity to transfer all data from SSs in PMP architecture. Any two (or more) SSs are not allowed to communicate directly. Transmission take place through two independent channels - downlink (DL) channel and uplink (UL) channel. The uplink channel is shared between all SSs while downlink channel is used only by the BS.

The standard defines both Time Division Duplex (TDD) and Frequency Division Duplex (FDD) for channel allocation. Both channels are time slotted and composed of frames. The TDD frame is composed of downlink and uplink sub-frames. The duration of each of these frames can be controlled by the BS whenever needed. The downlink channel is a broadcast channel and the BS broadcast data to all SS on the downlink channel. SSs accept only those packets which are destined for it, for more details about the uplink and downlink channel we refer the reader to references [1], [2].

### A. IEEE 802.16 Air Interface

The specified modulation scheme in the downlink and the uplink are binary phase shift keying (BPSK), quaternary PSK (QPSK), 16 and 64 quadrature amplitude modulation (QAM) to modulate bits to the complex constellation points. The FEC options are paired with the modulation schemes to form burst profiles. The PHY specifies seven combinations of modulation and coding rate, which can be allowed selectively to each subscriber, in both UL and DL. There are trade-offs between data rate and robustness, depending on the propagation conditions. Table 1 shows the combination of those modulation and coding rate.

TABLE I.  
MANDATORY CHANNEL CODING PER MODULATION [1]

Modulation	Uncoded block size	Coded Block size	Coding Rate $R$	RS code	CC code rate
BPSK	12	24	1/2	(12,12,0)	1/2
QPSK	24	48	1/2	(32,24,4)	2/3
QPSK	36	48	3/4	(48,36,2)	5/6
16QAM	48	96	1/2	(64,48,8)	2/3
16QAM	72	96	3/4	(80,72,4)	5/6
64QAM	96	144	2/3	(108,96,6)	3/4
64QAM	108	144	3/4	(120,108,6)	5/6

### B. IEEE 802.16 MAC

The IEEE 802.16 MAC layer is divided in three parts - Privacy sublayer (lower), MAC Common Part Sublayer (middle) and Convergence sublayer (upper). The core of the MAC layer is Common Part Sublayer (CPS). The MAC CPS is designed to support PMP and mesh network architecture and the MAC is connection oriented. Upon entering the network, each SS creates one more connections over which their data packets are transmitted to and from the BS. Each packet has to be associated with the connection at MAC level. This provides a way for bandwidth request, association of QoS and other traffic parameters and data transfer related actions. Each connection has a unique 16-bit connection identifier (CID) in downlink as well as in uplink direction.

The MAC Packet Data Unit (MPDU) is the data unit used to transfer data between MAC layers of BS and SS. The standard defines two types of MAC header - Generic MAC (GM) header and Bandwidth Request (BR) header. The generic header is used to transfer data or MAC messages while BR header is used to send bandwidth requests packets to BS. SSs send their bandwidth request in either bandwidth contention period or in allotted unicast uplink slots or piggybacked with data packets. The standard defines binary truncated exponential backoff algorithm for collision resolution in contention period and collisions only happens at the BS.

The standard defines a number of MAC management messages, which has to be transmitted between the SS and BS before actual data transfer. Any upcoming SS first synchronize itself with downlink channel to get Downlink Map (DL-MAP) and Uplink Map (UL-MAP) from the BS. DL-MAP and UL-MAP contains the information regarding downlink and uplink sub-frame, respectively. To setup a connection, each SS has to perform ranging, capacity negotiation, authentication, registration process in-sequence. Ranging process starts by sending Ranging Request (RANG-REQ) packets to BS in ranging contention slots. SSs do capability negotiation and registration process in-sequence after successful RANG-RSP. Registration is also done in request-response manner by sending Registration Request (REG-REQ) packet to BS and then the BS send REG-RSP packet back to the SS. Now any SS is ready to set up a connection with the BS and the connection formation is done in request-response manner.

The IEEE 802.16 standard supports four different flow

classes for QoS and the MAC supports an request-grant mechanism for data transmission in the uplink direction. These flows are associated with packets at MAC level. Each connection has a unique flow type associated with it. The IEEE 802.16 standard does not define any slot allocation criteria or scheduling algorithm for any type of service. A scheduling module is necessary to design UL-MAP to provide QoS for each SS and slot assignments for connections is done by BS and is included in the same UL-MAP. In particular, the WiMAX standard defines the following four types of services, each of which has different QoS requirements:

- 1) **Unsolicited Grant Services (UGS):** supports applications that generate fixed-size data packets periodically such as T1/E1 and VoIP without silence suppression. To support real-time needs of such applications and reduce overhead by the bandwidth request-grant process, the BS allocates a fixed amount of bandwidth to each of the flows in a static manner without receiving explicit requests from the SS.
- 2) **Real-Time Polling Services (rtPS):** support real-time traffic in which delay in an important QoS requirement. The amount of bandwidth required for this type of service is determined based on the required QoS performances, the channel quality, and the traffic arrival rates of the sources.
- 3) **Non-Real-Time Polling Service (nrtPS):** provides guarantees in terms of throughput only and is therefore suitable for mission critical data applications, such as File Transfer Protocol (FTP). The BS allows the SS to make periodic unicast grant requests, just like rtPS scheduling service, but the requests are issued at longer intervals.
- 4) **Best Effort Services:** provides no guarantees on delay or throughput and is used for Hypertext Transport Protocol (HTTP) and electronic mail (e-mail), for example. The bandwidth request for such applications is granted on space-available basis. The SS is allowed to use both contention-free and contention-based bandwidth requests, although contention-free is not granted when the system load is high.

While the concept of service flow is similar to a certain extent in both standards, IEEE 802.16e differs from IEEE 802.16-2004 in bandwidth grant services. In addition to the four data services listed above, IEEE

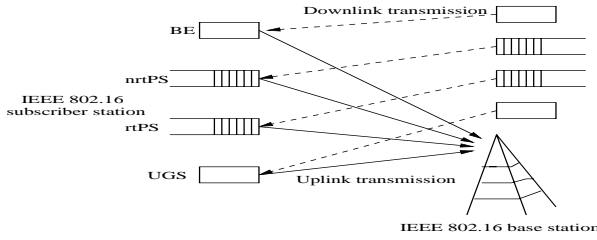


Figure 1. An IEEE 802.16 system model.

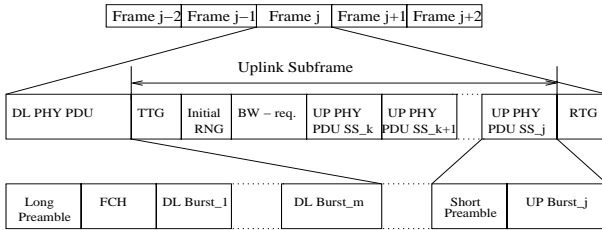


Figure 2. Frame structure in IEEE 802.16.

802.16e includes a new service known as extended rtPS which provides scheduling algorithm that builds on the efficiency of both UGS and rtPS. Similar to UGS, it is able to offer unsolicited unicast grants. However, the size of the bandwidth allocation is dynamic, unlike in UGS, in which the bandwidth allocation is fixed size.

Finally, to completely specify the QoS characteristics of a data service there are a number of mandatory QoS parameters that shall be included in the service flow definition when the scheduling service is enabled for it. Each scheduling service has a minimum number of associated parameters such as Minimum Reserved Traffic Rate<sup>1</sup> (MRTR), and Maximum Sustained Traffic Rate<sup>2</sup> (MSTR) etc.

### III. UPLINK SCHEDULING ALGORITHMS

The design of the uplink scheduling algorithm for IEEE 802.16 is more tricky than for the downlink since the UL does not have all information about the SSs such as the queue size. At the BS, the UL scheduling algorithm has to coordinate its decision with all the SSs whereas a DL algorithm is only concerned in communicating the decision locally to the BS. In this section, we will describe the evaluated UL scheduling algorithms considered in this article. Some of the UL scheduling algorithms is described in more detail whereas for the well-known algorithms such as WRR, EDF and WFQ, we refer the reader to references [6], [7], [13], [16], [17], [20].

<sup>1</sup>This parameter specifies the minimum rate reserved for the SS. The rate is usually expressed in bits per second and specifies the minimum amount of data to be transported on behalf of the SS when averaged over time. The MRTR rate will only be honored when sufficient data is available for scheduling.

<sup>2</sup>The parameter specifies the peak information rate of the SS. The value, expressed in bits per second, does not limit the instantaneous rate of the SS but it is used to police the SS to ensure that it conforms to the value specified, on average, over time.

#### A. Hybrid (EDF+WFQ)

This algorithm is a combination of the EDF and WFQ scheduling algorithms. This proposed "hybrid" algorithm allocates bandwidth among traffic classes based on the number of SSs and their MRTR in each class. The scheme utilizes EDF for SSs of the ertPS and rtPS whereas WFQ scheduling is used for SSs of the nrtPS and BE classes. The bandwidth is allocated in a fair manner and the overall bandwidth distribution is executed at the beginning of every frame while the plan EDF and WFQ algorithms are executed at the arrival of every packet. The following is the overall bandwidth allocation scheme adopted:

$$B_{ertPS,rtPS} = C \cdot \frac{\sum_{i \in ertPS,rtPS} MRTR_i}{\sum_{j=1}^n MRTR_j} \quad (1)$$

$$B_{nrtPS,BE} = C \cdot \frac{\sum_{i \in nrtPS,BE} MRTR_i}{\sum_{j=1}^n MRTR_j} \quad (2)$$

where  $C$  is the uplink channel capacity.

#### B. Cross-Layer Algorithm

The proposed cross-layer algorithm supports all service classes and provides isolating between the classes. Furthermore, it optimally calculates the number of slots in each frame such that the blocking probability of each class is minimized and the algorithm also integrates a proportional fairness scheme to SS channel quality information to allocate time-slots among connections of the same class.

1) *UGS*: For the UGS class, we consider the latency and the MRTR as QoS metrics. The priority function of the UGS connection  $m$  at time  $t$  is defined as

$$\theta_m^1(t) = \begin{cases} \frac{\beta_m^1}{\bar{\beta}_m^1} \frac{1}{\Delta t_m^1}, & \Delta t_m^1 > 0, \beta_m^1(t) \neq 0 \\ \infty, & \Delta t_m^1 = 0, \beta_m^1(t) \neq 0 \\ 0, & \beta_m^1(t) = 0 \end{cases}$$

where  $\beta_m^1$  is the  $m^{\text{th}}$  connection attainable bandwidth at time frame  $t$  and it reflects the channel quality between the BS and SS.  $\Delta t_m^1$  is defined as the latency bound of connection  $m$  and a new packet is time stamped by  $\Delta t_m^1$  which is decremented as long as the packet is queued. When a packet reaches its delay bound, its priority function becomes  $\infty$ . Furthermore, the factor  $\frac{\beta_m^1}{\bar{\beta}_m^1}$  defines the proportional fairness among users and  $\bar{\beta}_m^1$  is the average throughput for connection  $m$  at time  $t$  estimated over the window size  $1/\alpha$  and it is updated as follows

$$\bar{\beta}_m^1(t+1) = \begin{cases} \bar{\beta}_m^1(t)(1-\alpha), & m \notin C_1^*(t) \\ \bar{\beta}_m^1(t)(1-\alpha) + \alpha\beta_m^1(t), & m \in C_1^*(t) \end{cases}$$

where  $C_1^* \subseteq C_1(t)$  is the subset of connections that were selected to be served at the current time frame  $t$ .

TABLE II.  
MANDATORY QOS SERVICE FLOW PARAMETERS

Class of Service	Parameters	Possible applications
UGS	Maximum Sustained Traffic Rate Maximum Latency Tolerated Jitter Request/Transmission Policy Minimum Reserved Traffic Rate	ATM CBR; E17T1 over ATM; TDM Voice; VoIP without silence suppression
rtPS	Minimum Reserved Traffic Rate Maximum Sustained Traffic Rate Maximum Reserved Traffic Rate Request/Transmission Policy	Video telephony, VoD, AoD, Internet shopping
ertPS	Maximum Sustained Rate Traffic Priority Minimum Reserved Traffic Rate Maximum Latency Jitter Tolerance	VoIP with activity detection
nrtPS	Maximum Sustained Traffic Rate Minimum Reserved Traffic Rate Traffic Priority Request/Transmission Policy	High-speed file transfer, Multimedia messaging, E-commerce
BE	Maximum Sustained Traffic Rate Traffic Priority Request/Transmission Policy	Web-browsing, SMS, P2P file sharing

2) *rtPS*: For *rtPS* we consider the same QoS parameters as in *UGS* class, although the *rtPS* is not that sensitive to delay as *UGS*. The priority function of *rtPS* connection  $m$  at time  $t$  is defined as

$$\theta_m^2(t) = \begin{cases} \frac{\beta_m^2}{\beta_m} \frac{1}{\Delta t_m^2} \frac{\xi_m^2}{\xi_s^2}, & \Delta t_m^2 > 0, \beta_m^2(t) \neq 0, \xi_m^2 \neq 0 \\ \infty, & \Delta t_m^2 = 0, \beta_m^2(t) \neq 0, \xi_m^2 \neq 0 \\ 0, & \beta_m^2(t) = 0, \xi_m^2 = 0 \end{cases}$$

where  $\xi_m^2$  is the size of the queue of connection  $m$  and  $\xi_s^2 = \max_{m \in \mathcal{C}_2(t)} \xi_m^2$  for considering the amount of backlogged packets waiting for transmission. The set  $\mathcal{C}_2(t)$  is the set of all *rtPS* connections at time frame  $t$ , and  $\mathcal{C}_2^* \subseteq \mathcal{C}_2(t)$  is the set of served connections at time frame  $t$  based on the priority function  $\theta_m^2(t)$ .

3) *nrtPS*: For *nrtPS* we consider the minimum reserved bandwidth as priority metric and the priority function for SSs of *nrtPS* class is defined as follows:

$$\theta_m^3(t) = \frac{\beta_m^3}{\beta_m^3} \frac{\xi_m^3}{\xi_s^3}, \forall m \in \mathcal{C}_3(t), \quad (3)$$

where  $\xi_m^3$  is introduced to guarantee that no connection will be scheduled if there is no packets to transmit even if the channel quality is good.

4) *BE etc*: The priority of *BE* SSs depends only on the channel quality of the SS since the *BE* scheduling do not have any QoS requirements. The priority function for *BE* connection  $m$  at time  $t$  is

$$\theta_m^4(t) = \frac{\beta_m^4(t)}{\beta_m^4(t)}, \forall m \in \mathcal{C}_4(t), \quad (4)$$

where  $\beta_m^4(t)$  is the attainable bandwidth of connection  $m$  which captures the channel quality.

### C. Queuing Theoretic Scheduling Algorithm

We consider an uplink queuing theoretic (QT) scheduling algorithm which consists of a closely coupled scheduling algorithm and a CAC scheme [10]. The algorithm use a queuing model to satisfy the QoS requirements of the multi-class traffic. The QT algorithm uses utility functions to represent the level of users satisfaction on the perceived QoS for different service types. Hence, in this article we have neglected the implementation of the CAC since the focus is on the scheduling algorithm itself. The QT algorithm uses thresholds to limit the bandwidth allocated to SSs of each class. This is a unique way of limiting bandwidth allocation and ensuring that lower priority SSs do not starve. The utility of each SS is calculated at the start of the frame and bandwidth is allocated accordingly and the utility function for *BE* is defined as follows:

$$U_{BE}(b_i) = \begin{cases} 1, & \text{if } b_i(t) \leq 0 \\ 0, & \text{Otherwise} \end{cases}$$

For *rtPS*, *ertPS* and *nrtPS* connections, we use the modified sigmoid function [11] to obtain utility as a function of the packet-level performance measures. The utility functions can be expressed as follows:

$$U_{ertPS,erPS}(b_i) = 1 - \frac{1}{1 + \exp(-g_{rt}(\mathcal{K} - d_i^{req} - h_{rt}))} \quad (5)$$

$$U_{nrtPS}(b_i) = 1 - \frac{1}{1 + \exp(-g_{nrt}(\mathcal{X} - \tau_i^{req} - h_{nrt}))} \quad (6)$$

where  $\mathcal{K} = d(\bar{\gamma}, \lambda, b)$  and  $\mathcal{X} = \tau(\bar{\gamma}, \lambda, b)$  denote the average delay and transmission rate as functions of PDU arrival rate ( $\lambda$ ), average SINR ( $\bar{\gamma}$ ) when the allocated bandwidth is  $b$ .  $g_{rt}$ ,  $g_{nrt}$ ,  $h_{rt}$  and  $h_{nrt}$  are parameters of the sigmoid function and determine the steepness

(sensitively of the utility function to delay or throughput requirement) and the center of the utility function, respectively. The goal of the QT algorithm is to maximize the utility of all the SSs in the network and for this purpose an optimization problem is formulated (See [10] for details). An important part of the algorithm is determining bandwidth threshold of each traffic class and to be able to compare all the algorithms under the same constraints, we can calculate the thresholds as follows:

$$T_{class} = \frac{\sum_{i=1}^{n_{class}} MRTR_i}{\sum_{j=1}^n MRTR_j} \cdot C, \quad (7)$$

where  $\sum_{j=1}^n MRTR_j$  refers to the sum of MRTR of all the SSs in the network and  $\sum_{i=1}^{n_{class}} MRTR_i$  refers to the sum of SSs from different traffic classes.

#### IV. SIMULATION ENVIRONMENT AND NUMERICAL RESULTS

##### A. Simulation Environment

To assess the performance of the different scheduling algorithms we have used an ns-2 simulator where the MAC layer of SSs and the BS, including all procedure and functions for UL/DL data transmission and uplink bandwidth requests/grants. A detailed description of our design choices and implementation of the IEEE 802.11 standard can be found in [1] and [2]. According to the standard, the allocation start up time for OFDM PHY can either be in the start of the uplink subframe in the current frame or in the start of next subframe. The allocation start time is the reference point for the information in the UL-MAP message. In our simulations, the value for allocation start time is set such that all the allocation in the UL-MAP will start in the current frame after the last specified allocation in the DL-MAP.

Channel characteristics are simulated by using the empirical time-dispersive Stanford University Interim (SUI) channel models developed under the IEEE 802.16 Working Group. In our simulations we have considered SUI-A channels between the BS and SSs. We accounted for path loss and shadowing according to the following equation (valid for  $d > d_0$ ):

$$PL = 20 \log_{10}(4\pi d_0/\lambda) + 10\gamma \log_{10}(d/d_0) + X_f + X_h + s, \quad (8)$$

where,  $d$  is the distance between SS and the BS antennas in meters,  $d_0 = 100$  m, and  $s$  is a log normally distributed factor that accounts for the shadow fading with a standard deviation of value between 8.2 and 10.6 dB. The parameter  $\gamma$  is the path loss exponent. Correction factors  $X_f$  and  $X_h$  has been used to account for the operating frequency outside 2.5 GHz and the given terrain and SS antenna height above ground.

VoIP is modeled as an ON/OFF source with duration of exponentially distributed and packets are only generated during the ON period. Video traffic is generated by real MPEG4 traces [29]. The data traffic is modeled as a Web source and we considered a hybrid Lognormal/Pareto

distribution. The body of the distribution corresponding to an area of 0.88 is modeled as a Lognormal distribution with mean of 7247 bytes, and the tail is modeled as a Pareto distribution with a mean of 10558 bytes [30].

##### B. Performance Metrics

We have specified several metrics to assess the performance of the scheduling algorithms. The following metrics have been defined:

- 1) **Frame utilization:** the number of symbols utilized for data out all the symbols in the uplink sub-frame and can be defined as follows (metric is reported in percentage):

$$\mathcal{F} = \frac{\sum_{i=1}^n \bar{\omega}_i}{N_s} \times 100\% \quad (9)$$

where  $\bar{\omega}_i$  is the number of data symbols allocated to a SS,  $N_s$  is the total number of symbols in the uplink sub-frame and  $n$  is number of connections.

- 2) **Average Throughput:** is defined as the amount of data selected for transmission by a user per unit time.
- 3) **Average Delay:** is defined as the time between arrival of a packet in the queue to the departure of the packet from the queue and can be calculated for each SS according to:

$$\mathcal{D} = \frac{\sum_{i=1}^N (f_i - a_i)}{N}, \quad (10)$$

where  $f_i$  is the time packet  $i$  leaves the queue and  $a_i$  is the arrival time of packet  $i$  in the queue.  $N$  is the number of packets.

- 4) **Fairness:** In this article, we will use the Jain's fairness index which could be defined as follows [31]:

$$\mathcal{J} = \frac{(\sum_{i=1}^n \eta_i)^2}{n \cdot \sum_{i=1}^n (\eta_i)^2}. \quad (11)$$

Where we used the normalized throughput for  $\eta_i$ . The average throughput of a SS is normalized with respect to the MRTR of the SS, i.e.  $\bar{\eta}_i = \eta_i / MRTR$ .

##### C. Performance Results

1) *Effect of uplink burst preamble on user performance:* We investigate the different scheduling algorithms for both light and heavily loaded systems.

**Average Throughput:** Figures 3 and 4 show the obtained simulation results for both light and heavily loaded system. It is obvious that average throughput decreases with increasing number of SSs, this due to decreasing load per SS and increase in bandwidth wasted by uplink burst preambles. For less number of SSs, the Queueing theoretic algorithm is superior (for all classes) all other investigated scheduling algorithms due to it allocates at least one MRTR in every frame. Although under heavy load and

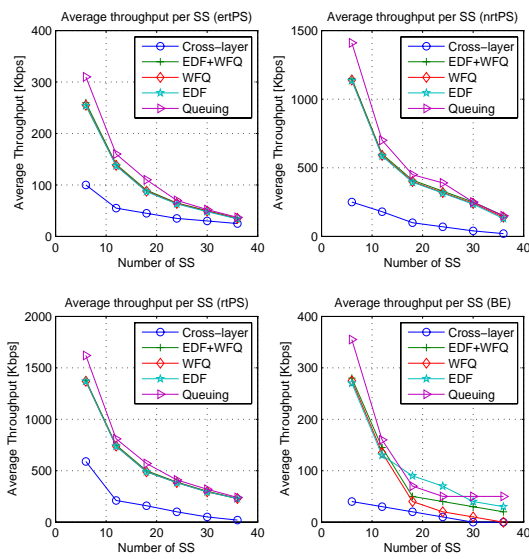


Figure 3. The effect of uplink burst preamble - Average throughput for light traffic load.

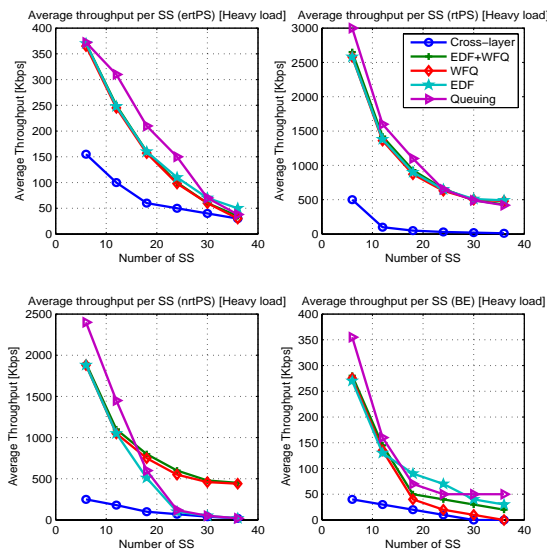


Figure 4. The effect of uplink burst preamble - Average throughput for heavy traffic load.

a large number of SSs its throughput performance is less than the hybrid algorithm (EDF+WFQ). This can easily be explained by the fact of large overhead as it selects the maximum number of SSs in a frame. SSs of the BE class get assigned bandwidth equivalent to the size of one packet (100 bytes). Due to large packet size, the average throughput of BE SSs under heavy load and large number of SSs stays at 50 Kbps. It is also noted that the average throughput of SSs of the nrtPS class is low with a large number of SSs and under heavier load. This due to the minimum allocated bandwidth per frame by the Queuing theoretic algorithm is smaller than the packet size of FTP

traffic. With a large number of SSs, the ertPS and rtPS SSs will be assigned higher priority by the Queuing theoretic algorithm due to its high average delay. This will in turn result in very few transmission opportunities for the nrtPS SS and therefore their MRTR will not be satisfied. The cross layer algorithm indicates the lowest throughput of all algorithms, this due to it selects only one SS in a frame. When the number of SSs increases, the lower priority SSs (nrtPS and BE classes) will allocated little bandwidth. The WFQ algorithm have a similar throughput performance as the EDF and hybrid scheduler.

**Average Delay:** In figure5 the average delay is plotted as a function of number of SSs. As expected, the average delay will increase with larger number of SSs, due to increasing overhead of uplink burst preamble. The cross-layer algorithm do not experience any significant increase in delay when we increase the traffic load. This could be explained by that the algorithm only selects one SS in a frame and will result in a large backlog data. Backlogged packets will miss their deadline and will be dropped.

**Fairness:** Figures 7 and 8 show the fairness for different evaluated algorithms. For the ertPS class, it is observed that the fairness for cross-layer, WFQ and Queuing theoretic algorithms decreases when the number of SSs increases. In the case of cross-layer algorithm, only one SS selected and hence the difference in minimum and maximum throughput in this class will be rather high. For QT-algorithm, some SSs will receive a large portion of the bandwidth (more than their MRTR) and some will not. This is the main reason for the decrease in fairness among SSs of the rtPS class. The WFQ algorithm indicates a low fairness among ertPS SSs due to the bursty nature of VoIP traffic. In [32], Shi and Sethu showed that time-stamp based WFQ schedulers have a low fairness among users for bursty traffic. For the QT-algorithm, some SSs will receive a large portion of bandwidth and some will not. This could easily be explained by that the utility function of rtPS SSs do not take the average throughput into account which results in fluctuation of intra-class fairness. The intra-class fairness of the rtP class under the cross-layer algorithm depends on the amount of transmitted traffic by the selected SS, e.g., when the number of SS is 12 and 18.

2) *The effect of frame length:* We also study the performance of the scheduling algorithms when different frame lengths is employed. The considered frame lengths are: 2.5ms, 4ms, 5ms, 8ms, 10, ms, 12.5ms and 20ms. With larger frame sizes, subscriber stations can send more data due to more symbols available in the frame. For the largest frame size (20ms) and a symbol duration of 12.5μs, the total number of symbols available for the uplink subframe is 800, which could be compared to 100 symbols when a frame size of 2.5ms is considered. As before, we will consider both light and heavy loaded. Under light load, each rtPS SS will send traffic at a rate of 1000 Kbps and the nrtPS SS will send traffic at a rate of 500 Kbps. For the scenario of heavy load, each rtPS SS sends traffic at a rate of 2000 Kbps and nrtPS SS at

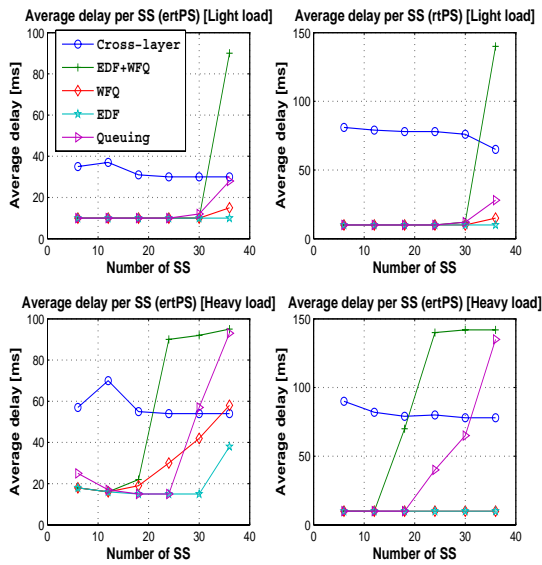


Figure 5. The effect of uplink burst preamble - Average delay for both light and heavy traffic load.

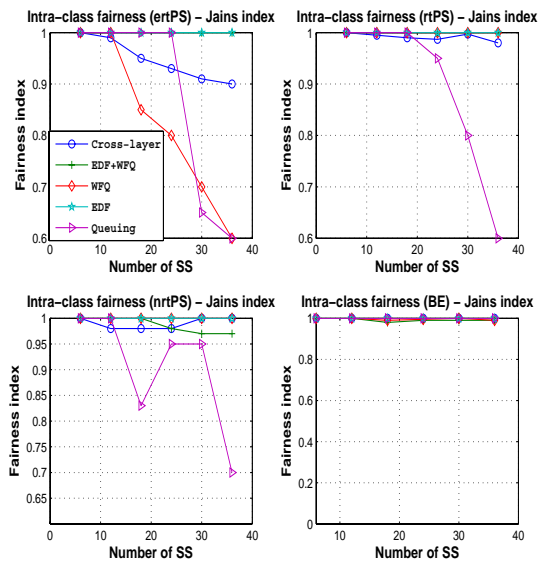


Figure 7. The effect of uplink burst preamble - Intra-class fairness for heavy traffic load.

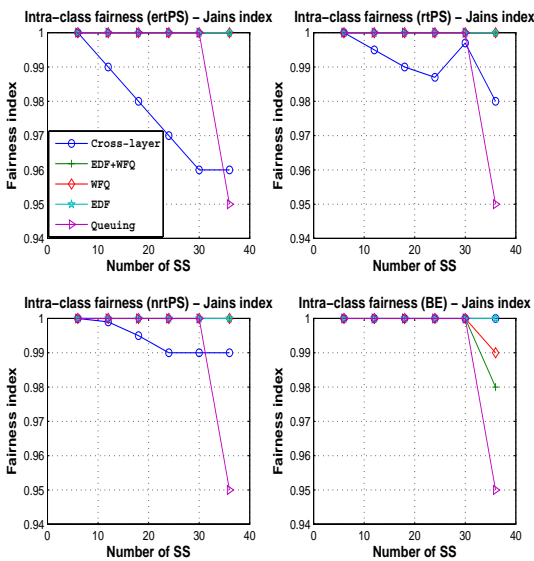


Figure 6. The effect of uplink burst preamble - Intra-class fairness for light traffic load.

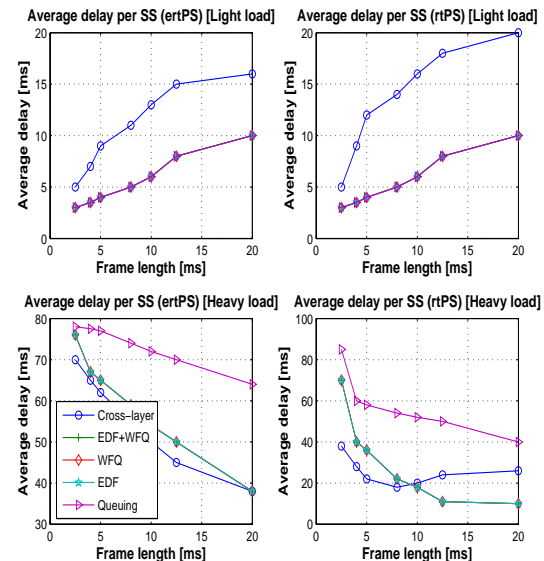


Figure 8. The effect of frame lengths - Average delay for light and heavy traffic load.

a rate of 1000 Kbps.

**Average Delay:** Figure 8 show the average delay of SSs for both the ertPS and rtPS class for different frame sizes under light load. The average delay increases with increase frame size. This can be explained by the fact that packets spend longer time in the queue. When considering heavily load, the average delay of SSs from both ertPS and rtPS will decrease with increasing frame size. This is mainly due to more packets being flushed out of the queue of the SSs. The cross-layer algorithm indicates an increase in average delay for the rtPS class as the packets wait a longer time in the queue due to larger frame size.

The average delay of SSs of the ertPS class under TQ algorithm is higher than indicated by other algorithms, this is mainly due to the fact that QT tends to satisfy all the SSs MRTR requirements and allocate the residual bandwidth according to the utility of the SSs. Depending on the channel quality, some SSs will not be allocated bandwidth any further, thus their packet spend a longer time in the queue.

**Frame Utilization:** Considering light load and small file size, the cross-layer algorithm has higher frame utilization than both WRR and QT as shown in Fig. 9. This can mainly be explained by the fact that WRR

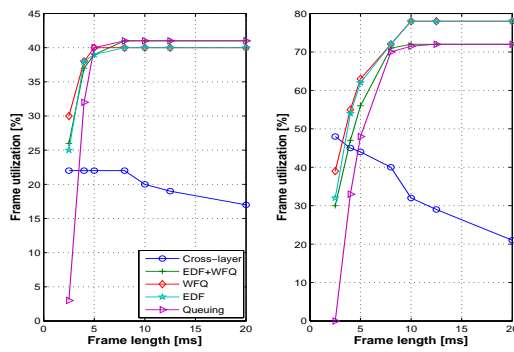


Figure 9. Frame utilization under light load and heavy load.

and QT selects the maximum number of SSs in a frame which results in the largest overhead. When the frame size increases, the frame utilization of WRR and QT rapidly increases and reaches the same performance as the other schemes. When considering smaller frame size and heavy load as in Fig. 10, the cross-layer algorithm indicates the highest frame utilization as the overhead from selecting multiple SSs by the other algorithms are significant. When frame size increases, the QT algorithm increases the frame utilization while the cross-layer algorithm performance decreases due to the amount of data a single SS has to send remains the same.

V. CONCLUDING REMARKS

In this paper, we have proposed two scheduling algorithms and evaluated several other legacy scheduling algorithms for the uplink in IEEE 802.16 network aiming at satisfying QoS requirements of the multi-class traffic. The algorithms are evaluated under a different mix of traffic and with respect to the major characterization of IEEE 802.16 MAC layer such as bandwidth request mechanisms, frame size and the uplink burst preamble. The first algorithm that is proposed is a combination of earliest deadline first and weighted fair queuing. The algorithm provides a more fair distribution of bandwidth among the SSs than most of the evaluated scheduling algorithms. The second algorithm is an opportunistic cross-layer scheduling scheme which support all QoS classes in the IEEE 802.16 standard. The proposed scheduling algorithm is optimal, fair and capable of isolating the flow classes.

Simulation reveals that legacy scheduling algorithms such as EDF, WFQ and EDF+WFQ do not explicitly consider all required QoS parameters of the traffic classes in IEEE 802.16. This is not sufficient since scheduling classes have multiple QoS parameters such as rtPS requiring delay, packet loss and throughput guarantee. Cross-layer and queuing theory scheduling algorithm is more suitable since they include the maximum latency, MRTR and the channel quality in the priority functions. Although, they have also some drawbacks such as implementation complexity (especially the queuing theory scheduling algorithm).

ACKNOWLEDGMENT

The authors wish to thank the anonymous reviewers for their careful reviews and constructive comments that helped improve the article.

REFERENCES

- [1] *IEEE 802.16-2004, IEEE standard for local and metropolitan area networks part 16: air interface for fixed broadband wireless access systems*, IEEE, Apr. 2005.
- [2] *IEEE P802.16/Cor1/D2, Corrigendum to IEEE standard for local and metropolitan area networks part 16: air interface for fixed broadband wireless access systems*, IEEE, Apr. 2005.
- [3] G. Chu, D. Wang, and S. Mei, "A QoS architecture for the MAC protocol of IEEE 802.16 BWA system," in *Proc. IEEE Conference on Communications, Circuits and Systems and West Sino Expositions*, June 2002.
- [4] K. Wongthavarawat and A. Ganz, "Packet scheduling for QoS support in IEEE 802.16 broadband wireless access systems," *International Journal of Communication Systems*, vol. 16, issue 1, pp. 81-96, Feb. 2003.
- [5] J. Lin and H. Sirisena, "Quality of service scheduling in IEEE 802.16 broadband wireless networks," in *Proc. First International Conference on Industrial and Information Systems*, Aug. 2006.
- [6] C. Liu and J. Layland, "Scheduling algorithms for multiprogramming in a hard real-time environment," *Journal of ACM*, vol. 20, pp. 46-61, 1973.
- [7] D. Ferrari and D. Verma, "A scheme for real-time channel establishment in wide-area networks," *IEEE J. Select Areas Commun.*, vol. 8, pp. 368-379, Apr. 1990.
- [8] N. Ruangchaijatupon, L. Wang, and Y. Ji, "A study on the performance of scheduling schemes for broadband wireless access networks," in *Proc. International Symposium on Communications and Information Technology (ISCI'06)*, Oct. 2006.
- [9] Q. Liu, W. Wang and G. B. Giannakis, "Cross-layer scheduler design with QoS support for wireless access networks," *Proc. International Conference on Quality of Service in Heterogeneous Wired/Wireless Networks (QSHINE'05)*, Aug. 2005.
- [10] D. Niyato and E. Hossain, "A queuing-theoretic optimization-based model for radio resource management in IEEE 802.16 broadband wireless networks," *IEEE Trans. on Computers*, vol. 55, no. 11, pp. 1473-1488, Nov. 2006.
- [11] M. Xiao, N. B. Shroff and E. K. P. Chong, "Utility-based power control in cellular wireless systems," *Proc. IEEE INFOCOM'01*, 2001.
- [12] D. Niyato and E. Hossain, "Queue-aware uplink bandwidth allocation and rate control for polling service in IEEE 802.16 broadband wireless networks," *IEEE Trans. on Mobile Computing*, vol. 5, pp. 668-679, June 2006.
- [13] J. C. R. Bennett and H. Zhang, "WF<sup>2</sup>Q: Worst-case fair weighted fair queuing," in *Proc. IEEE INFOCOM'96*, 1996.
- [14] S. Kim and I. Yeom, "TCP-aware uplink scheduling for IEEE 802.16," *IEEE Commun. Letters*, pp. 146-148, Feb. 2007.
- [15] M. Settembre, M. Puleri, S. Garritano, P. Testa, R. Albanese, M. Mancini and V. Lo Curto, "Performance analysis of an efficient packet-based IEEE 802.16 MAC supporting adaptive modulation and coding," in *Proc. International Symposium on Computer Networks (ISCN'06)*, June 2006.
- [16] C. Cicconetti, A. Erta, L. Lenzini and E. Mingozzi, "Performance evaluation of the IEEE 802.16 MAC for QoS support," *IEEE Trans. on Mobile Computing*, vol. 6, no. 1, Jan. 2007.

- [17] M. Katevenis, S. Sidiropoulos and C. Courcoubetis, "Weighted round-robin cell multiplexing in a general-purpose ATM switch chip," *IEEE J. Sel. Areas Commun.*, vol. 9, pp. 1265-1279, Oct. 1991.
- [18] H. Lee, T. Kwon and D.-H. Cho, "An efficient uplink scheduling algorithm for VoIP services in IEEE 802.16 BWA systems," in *Proc. Vehicular Technology Conference (VTC'04-fall)*, Sept. 2004.
- [19] C. Eklund, R. B. Marks, K. L. Stanwood, and S. Wang, "IEEE Standard 802.16: A technical overview of the wirelessMAN air interface for broadband wireless access," *IEEE Commun. Magazine*, vol. 40, no. 6, pp. 98-107, June 2002.
- [20] P. Bhattacharya and A. Ephremides, "Optimal scheduling with strict deadlines," *IEEE Trans. Automatic Control*, vol. 34, pp. 721-728, 1989.
- [21] S. Panwar, D. Towsley and J. Wolff, "Optimal scheduling policies for a class of queues with customer deadlines to the beginning of service," *J. Ass. Comput. Mach.*, vol. 35, pp. 832-844, 1988.
- [22] L. Georgiadis, R. Guerin and A Parekh, "Optimal multiplexing on a single link: delay and buffer requirements," *IEEE/ACM Trans. Info. Theory*, vol. 43, pp. 1518-1535, 1997.
- [23] L. Georgiadis, R. Guerin and V. Peris, "Efficient network QoS provisioning based on per node traffic shaping," *IEEE/ACM Trans. Networking*, vol. 4, pp. 482-501, 1996.
- [24] J. Chen, W. Jiao, and Q. Guo, "Providing integrated QoS control for IEEE 802.16 broadband wireless access systems," *IEEE Vehicular Technology Conference (VTC'05-fall)*, Sept. 2005.
- [25] J. Xu and R. Lipton, "On fundamental tradeoffs between delay bounds and computational complexity in packet scheduling algorithms," in *Proc. SIGCOMM'02*, Oct. 2002.
- [26] H. Fattah and C. Leung, "An overview of scheduling algorithms in wireless multimedia networks," *Proceedings IEEE Wireless Communication*, pp. 76-83, Oct. 2002.
- [27] K. Chen, A. Liu and L. Lee, "A multiprocessor real-time process scheduling method," in *Proc. of 5th International Symposium on Multimedia Software Engineering*, Dec. 2003.
- [28] The network simulator - NS-2,  
<http://www.isi.edu/nsnam/ns>
- [29] P. Barford *et al.*, "Changes in web client access patterns: Characteristics and caching implications," Technical report 1998-2003, Boston University, 2003.
- [30] P. Seeling, M. Reisslein, and B. Kulapla, "Network performance evaluation using frame size and quality traces of single-layer and two-layer video: A tutorial," *IEEE Commun. Mag.*, vol. 6, no. 2, pp. 58-78, 2004.
- [31] R. Jain, D. M. Chiun, and W. Hawe, "A quantitative measure of fairness and discrimination of resource allocation in shared systems," *Digital Equipment Corporation*, DEC-TR-301, 1984.
- [32] H. Shi, H. Sethu, and S. S. Kanhere, "An evaluation of fair packet schedulers using a novel measure of instantaneous fairness," *Computer Communications* 28(17): 1925-1937 (2005)

**Gang Wang** received his MSc degree in computer science from Rutgers, the State University of New Jersey, USA, in 2000.

He is currently working as Expert at Philips Research, Eindhoven, The Netherlands. Before joining Philips Research, he was a consultant for TietoEnator, Stockholm, Sweden. His research interests include wireless communication and signal processing.

**Mikael Gidlund** received his MSc in electrical engineering from Mid Sweden University, Sundsvall, Sweden in 1999 and his Lic. Tech degree in radio communication systems from Royal Institute of Technology (KTH), Stockholm, Sweden in 2004 and his Ph.D. degree in electrical engineering from Mid Sweden University, Sundsvall, Sweden in 2005.

He is currently working as research scientist at Nera Networks AS, Bergen, Norway. Between February 2006 and August 2007 he was working as senior research engineer and project manager at Acreo AB, Sweden. During February-July 2005 he was visiting researcher at the Dept. of Informatics, University of Bergen, Norway. His current research interests are in communication theory, signal processing and multimedia transmission for wireless communication, with particular emphasis on error control techniques, radio resource management and multimedia distribution.

Dr Gidlund has served as member of technical program committee at the IEEE Information Theory Workshop (ITW'07), IEEE PIMRC'08, IEEE VTC'09-spring, IEEE ISWCS'09. He is a member of IEEE and EuMA.