Packet Scheduling with QoS and Fairness for Downlink Traffic in WiMAX Networks

Wei Nie*, Houjun Wang* and Jong Hyuk Park**

Abstract—The IEEE 802.16 standard is supposed to provide a wide-range broadband wireless service, but it leaves the implementation of the wireless resource scheduler as an open issue. We have studied the scheduling problem and propose a two level scheduling (TLS) scheme with support for quality of service and fairness guarantees for downlink traffic in a WiMAX network. A central controller Base Station has a number of users, and each mobile subscriber station has different channel conditions. The same mobile subscriber station may have different service requirements at different times in the WiMAX network. Based on OPNET simulation, the results show our scheduling algorithm can increase the network throughput, maintain relative fairness, and lower delay over the round robin and weighted round robin algorithms.

Keywords—IEEE.802.16, Scheduling, Round Robin, Weighted Proportional, Fairness

1. INTRODUCTION

Recently demand for high-speed internet and multimedia service for residential and business customers has increased greatly. IEEE 802.16 (also known as Worldwide Interoperability for Microware Access System) is a new standard based on Broadband Wireless Access [1]. It is a set of telecommunication technology standards aimed at providing wireless access over long distances in a full-mobile, cellular-type access.

A WiMAX System covers a metropolitan area of several kilometers and is also called WirelessMAN. A basic WiMAX network consists of a base station (BS) and multiple subscriber stations (SSs). Theoretically, a WiMAX base station can provide broadband wireless access in ranges up to 50Kms for mobile stations with a maximum data rate of up to 70 Mbps. An important principle of WiMAX is that it is connection oriented. This means that an SS must register with the base station before it can start to send or receive data. During the registration process, an SS negotiates the initial Quality of Service (QoS) requirements with the BS. These requirements can be changed later, and a new connection may also be established on demand. The QoS requirements may be either per-connection (GPC) or per-subscriber station (GPSS). In this paper, we limit the discussion to GPC. The Base Station does the scheduling for downlink direction and provides the QoS guarantees in the WiMAX network. An algorithm at the Base Station has to translate the QoS requirements of the SSs into the appropriate number of slots. There have been few scheduling architectures reported in the literature.

Manuscript received October 1, 2010; first revision November 29, 2010; accepted January 7, 2011.
Corresponding Author: Wei Nie
* School of Automation Engineering, University of Electronic Science and Technology of China (weiuestc@gmail.com, hjwang@uestc.edu.cn)
** Department of Computer Science and Engineering, Seoul National University of Science and Technology, Korea (jhpark1@snut.ac.kr)

In this paper we propose a packet scheduling with QoS and fairness guarantees for downlink traffic in WiMAX networks. The remainder of this paper is organized as follows: In Section II we summarize some related work in the area of IEEE 802.16. In Section III we describe the downlink traffic scheduling schemes at the BS, which are based on two level scheduling with the first level ensuring QoS and the second level providing fairness. In Section IV we present the preliminary simulation results and compare them with the round robin and weighted round robin algorithms. Finally, Section V concludes the discussion and presents our future work.

2. RELATED WORK

The main purpose of the downlink scheduler is bandwidth allocation. The IEEE 802.16 standard defines a frame structure as depicted in Fig 1. The vertical axis in this figure is frequency or subcarriers and the horizontal axis is time. The time is divided into frames (typically 5 ms duration) [2].Each frame consists of downlink (DL) and uplink (UL) subframes. A preamble is used for time synchronization. The downlink map (DL-MAP) and uplink map (UL-MAP) define the burst-start time and burst-end time, modulation types and forward error control (FEC) for each SS. The MAP's lengths and usable subcarriers are defined by the Frame Control Header (FCH). The SS allocation is in terms of bursts. Since the channel state condition keeps changing over time because of the nature of wireless media, WiMAX supports adaptive modulation and coding; in other words, the modulation and coding can be changed adaptively depending on the channel condition[3]. Both the MS or BS can do the estimation, and then the BS decides the most efficient modulation and coding scheme. A Channel Quality Indicator (CQI) is used to pass the channel state condition information. Some parameters and attributes are needed in designing a WiMAX scheduler. Moreover, in the downlink direction, the IEEE 802.16e standard requires that all DL data bursts be rectangular.

In fact, the two-dimensional rectangular mapping problem is a variation of the bin packing problem, in which one is given bins to be filled with objects. The bin packing problem [4,5], or the rectangular packing problem[6,7], is also an important issue to be considered. It is known to

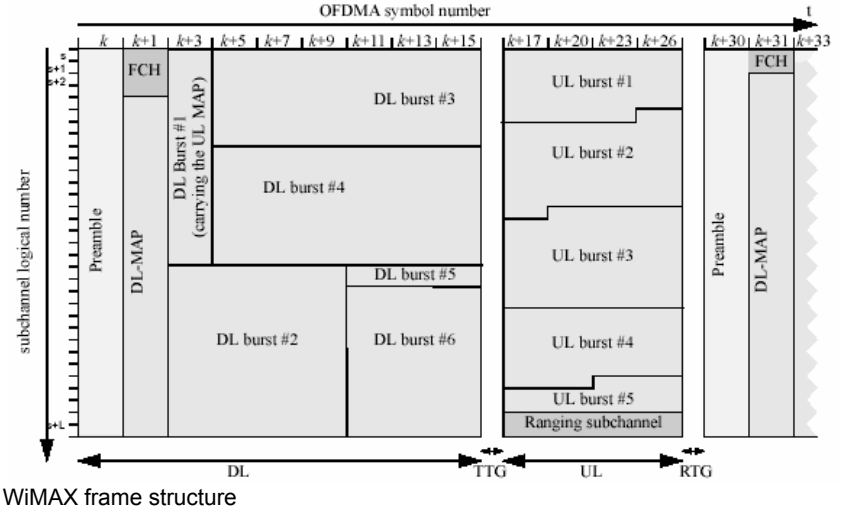

Fig. 1. WiMAX frame structure

be an NP hard problem [8]. However, if we can improve the efficiency of such packing, we can improve the bandwidth utilization of the WiMAX system. The packet schedulers operating at the MAC layer are very important for QoS delivery. The IEEE 802.16 standard does not specify the scheduling algorithm to be used. It is up to vendors to implement an algorithm based on their network traffic. Vendors and operators have their choice among many existing scheduling techniques or they can develop their own scheduling algorithms.

Round-Robin scheduling algorithms [9] are the simplest scheduling algorithms designed especially for time sharing systems. The RR algorithm can be considered the very first simple RR and it fairly assigns allocation one by one to all connections. With packet-based allocation, stations with larger packets have an unfair advantage. Moreover, RR may be non-work conserving in the sense that the allocation is still made for connections that may have nothing to transmit; since RR cannot assure QoS for different service classes.

The Weighted Round Robin algorithm has been applied for WiMAX scheduling [10], A WRR algorithm assigns weight to each SS and the bandwidth is then allocated according to the weights. The weights can be used to adjust for the throughput and delay requirements. The WRR is an extension of the Round Robin algorithm. It is a work-conserving algorithm in that it will continue allocating bandwidth to the SSs as long as they have backlogged packets.

3. DOWN LINK SCHEDULING ALGORITHMS

3.1 Overview of Downlink Traffic Scheduling Scheme at BS

The Downlink (DL) Scheduler in the BS distributes the entire downlink bandwidth among all downlink connections. When the uplink sub-frame ends, the BS first broadcasts the DL-MAP and UL-MAP, then the RANG-RSP messages, the REGREP messages, and the CONN-RSP messages, and lastly it starts sending downlink data packets to SSs. The downlink channel is an Always Broadcast channel. The BS Downstream Generator sends specific amounts of data from each downlink connection according to the output of the Downlink Scheduler.

Supporting QoS is a fundamental part of the WiMAX MAC layer design. In MAC connection oriented architecture, data is transmitted between BSs and SSs in the context of a connection. Each connection is identified in the (PDU) by a connection identifier (CID) which also provides

Table 1. Summary of QoS categories

QoS category	Applications	QoS specifications
UGS Unsolicited grant service	VoIP	Maximum sustained rate Maximum latency tolerance Jitter tolerance
rtPS Real-time polling service	Streaming audio or video	Minimum reserved rate Maximum sustained rate Maximum latency tolerance Traffic priority
ErtPS Extended real-time polling service	Voice with activity detection (VoIP)	Minimum reserved rate Maximum sustained rate Maximum latency tolerance Jitter tolerance Traffic priority
nrtPS Non-real-time polling service	File Transfer Protocol (FTP)	Minimum reserved rate Maximum sustained rate Traffic priority
BE Best effort service	Data transfer, Web browsing, etc.	Maximum sustained rate Traffic priority

a mapping to a service flow identifier (SFID). The SFID is an important concept in the MAC layer standard; it provides a mapping to the QoS parameters for a particular data entity.

WiMAX networks support different types of traffic. IEEE 802.16 defines the following five types of service flow with distinct QoS requirements: UGS Ertps, Rtps, Nrtps, and BE. A summary of the QoS categories can be found in Table 1[11]. A WiMAX multi-service environment is complex because of the various packet streams it needs to serve and their different QoS requirements and traffic behaviors. Because of the distinct QoS characteristics of each service type, a single scheduling algorithm may not be sufficient. In order to guarantee different QoS requirements for each queue and improve the throughput for downlink traffic in WiMAX networks, we propose packet scheduling with QoS and a fairness scheduling scheme. The service flows are scheduled by different algorithms.

3.2 QoS Priority and Fairness Scheduling Design

In the first level, Strict-Priority packets are first classified by the scheduler according to their QoS class: UGS > Ertps > Rtps > Nrtps > BE. Then they are placed into different priority queues. It services the highest priority queue until it is empty, and then moves to the next highest priority queue.

In the second level, we use a fairness-oriented scheduling scheme for different service flows. The BS provides fixed-size data grants at periodic intervals for UGS services, Adaptive Proportional Fairness (APF) scheduling for rtPS and ertPS services, and Proportional Fairness （PF） scheduling for Nrtps and BE services.

Proportional Fairness (PF): The goals of this packet scheduling scheme are to enhance the system throughput as well as to provide fairness among the queues. Although PF is simple and efficient, it cannot guarantee any QoS requirements, such as delay and jitter, due to the fact that it was originally designed for saturated queues with non-real-time data services, so we use this algorithm for Nrtps and BE services.

For time t, and user k, the channel quality is $ChCond_k(t)$. This is directly proportional with the signal which the SS receives and determines the maximal transmission rate. $ChCond_k(t)$ is calculated based on modulation, coding and repeat times. For time t, and user k, the historical throughput is $Th_k(t)$. The user's priority is calculated by the following formula:

$$\Pr iority_k(t) = \frac{ChCond_k(t)}{1 + Th_k(t)} \tag{1}$$

The user with the highest priority will be scheduled with the highest probability. The Adaptive Proportional Fairness (APF) scheduling scheme aims at extending PF [3] scheduling to real time services application and provides various QoS requirements. The scheduling scheme is based on the Grant Per Type-of Service (GPTS) principle, which aims at differentiating the delay performance of each queue. A novel priority function is devised for all the QoS guarantees, including rtPS and ertPS, for allocating time slots on the queues with the highest priority value. At the time interval t, the priority function for queue i is defined below in formula 2.Ri (t) is the real-time service minimum rate requirement, C_i (t) is the number of connections at the ith queue, and $ChCond_k(t)$ is the channel quality at time t. Channel quality time determines the transmis-

sion capacity. We calculate the APF priority as follows:

$$\Pr iority_k(t) = \frac{ChCond_k(t)}{1 + R_i(t)Th_k(t)/C_i} \tag{2}$$

Each queue corresponds to one QoS requirement class. We schedule the user according to its priority. For the fairness of my algorithm, we can calculate priority in formulas (1) and (2); both equations schedule an SS in a given round with a given throughput. In the next round, it will be less probable to schedule the same SS. The scheduling should work along these lines, for a given channel condition, the higher the throughput, the lower the priority. Similarly, for a given throughput, the higher the channel condition, the higher the priority.

The scheduler services a different queue according to the SS's priority. The priority of the real time services application is higher than the non-real time services application. So if the real time services application queue is empty, the scheduler will move to the non-real time services application queue. When the network is congested, and more real time services applications arrive, the scheduler will stop scheduling the non-real time services application and schedule the real time services application. When the network traffic is light, the scheduler will resume scheduling the non-real time services application queue.

3.3 Specific implementation steps

At the beginning of each frame, the system checks all available bandwidth BW and the BS provides fixed bandwidth BWUGS at periodic intervals for UGS services. BWrtPS (i∈ rtPS) is the bandwidth required by the rtps service for a given time. And the SS will be scheduled according to the priority.

If $BW - BW_{UGS} < \sum_{i \in rtPS} BW_{rtPS}$, we allocate the bandwidth by priority as shown above then the connection is assigned bandwidth. The remaining packets which are not scheduled will be discarded.

If $\sum_{i \in rtPS} BW_{rtPS} < BW - BW_{UGS}$, we guarantee the bandwidth of rtps packets, and then we try to allocate the greater bandwidth to nrtps and BE service.

4. SIMULATION AND RESULTS

4.1 Simulation Parameter

In this section, we analyze the performance of a two-level scheduling scheme and compare it to round robin and weighted round robin algorithms for Downlink Traffic in a WiMAX network. The simulations will be implemented in an OPNET [12] environment with the following parameters for the physical and MAC layer of a WiMAX network. The simulation parameters are as shown in table 2.

4.2 Traffic Model

We have implemented four different traffic sources, one for each of the traffic classes. VoIP

Table 2. Simulation parameters

Parameter Type	Parameter Value
Base frequency	2.5 GHz
Duplexing Mode	TDD
System bandwidth	5 Mbps
DL/UL ratio	2:1 (29:18 OFDM symbols)
Frame length	5 ms
Cyclic prefix duration	11.42 usec
Basic symbol	91.43 usec
Fast Fourier Transform Size	1024
PHY	OFDMA
DL permutation zone	PUSC
MAC PDU size	Variable length
Fragmentation	Enable
ARQ and Packing	Disable
DL-UL MAPS	Variable
Inter-arrival time between Video frames	120ms

traffic is modeled for SSs of UGS or ertPS, video streaming for SSs of rtPS, FTP for SSs of nrtPS class and HTTP for SSs of BE class.

Video Conference Traffic Model (Class 2) [2]

The audio and video components of video conferencing have different bandwidth requirements. The audio component requires between 16 and 64 Kbps and the video component requires between 320 Kbps and 1 Mbps. We implement a typical business-quality video conference which runs at 384 Kbps and can deliver TV-quality video at 25 to 30 frames per second. The traffic model of video conferencing is shown in table 3.

Web Browsing (HTTP) Traffic Model [2]

An HTTP traffic model is used for the BE class. Each webpage might consist of a number of web objects including embedded images, style sheets, executable java applets, plug-ins, and the page itself. The time between visiting two pages is referred to as the reading time and includes time for the user to read all or part of the page. The session is cut into ON/OFF periods representing webpage downloads and the intermediate reading times, where the webpage downloads are referred to as packet calls.

FTP Traffic Model (Class 5) [2]

We have implemented an FTP traffic generator with a constant packet size of 150 bytes. According to packet size distributions, 76% of the files are transferred using an MTU of 1500 bytes

Table 3. Video streaming parameters

Inter-arrival time between Video frames	120ms
Video packet size	Geometric(mean=200bytes)
Minimum Reserved Traffic Rate	64kbps
Maximum Sustained Traffic Rate	400kbps
Maximum latency	180ms
Tolerated packet loss	5%
Average traffic rate	220kps

and 24% of the files are transferred using an MTU of 576 bytes. These two packet sizes also include a 40 byte IP packet header.

4.3 Simulation scenario and result

We simulated a simple WiMAX network using OPNET Modeler v.16. The network consists of one BS with eighteen SS's in each cell as illustrated in Fig. 2. The source of traffic is an application server that serves five types of applications—one for each type of traffic. The assumption is that each SS carries the traffic from one user only, and each user utilizes only one type of application at a time. A simple priority queue data structure is implemented, which prioritizes different types of services based on strict priority rules. UGS is given the highest priority and BE the lowest.

4.3.1 Throughput comparisons

Fig. 3 shows the average throughput of BE service flows in the downlink. We can conclude that the BE service flow has better throughput in our algorithm than in the RR algorithm. Moreover, the throughput of our two-level scheduling scheme is higher than that of the Round Robin algorithm.

4.3.2 Queuing Delay

Queuing delay is shown in Fig. 4, which is the time between the arrival and departure of a packet from the queue. The value reported is an average calculated in milliseconds (ms).

The average delay is directly proportional to the number of SS's because of mutual competition for bandwidth. A videophone is one kind of real time service. Videophones typically use two way communications and have data rates of 32 - 384 kbps. With videophones, end-to-end delays less than 400 ms are acceptable, which means scheduled delays less than 120 ms will be accepted. From picture 4, compared with the WRR algorithm, the scheduler delay of the TLS algorithm is at most 90ms. This will meet the delay requirement and guarantee the QoS of real

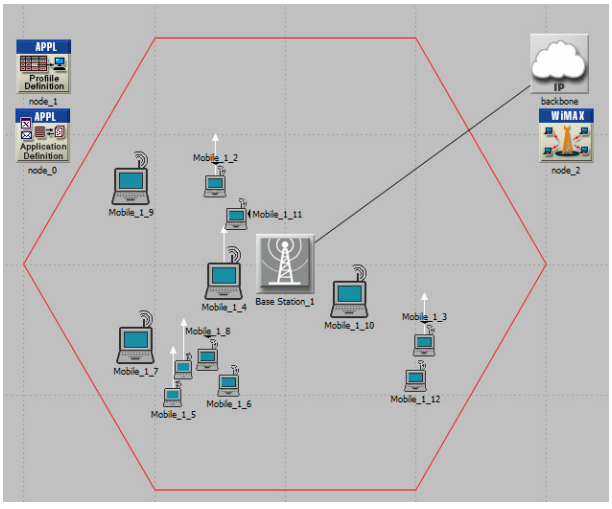

Fig. 2. Simulation scenario

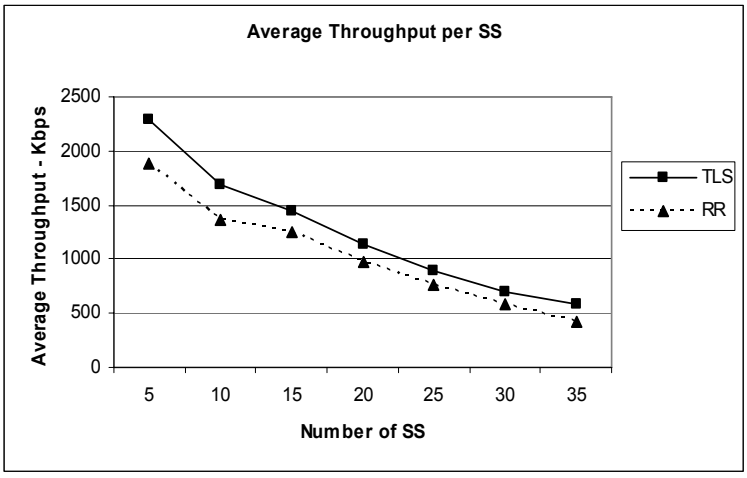

Fig. 3. Average Throughput

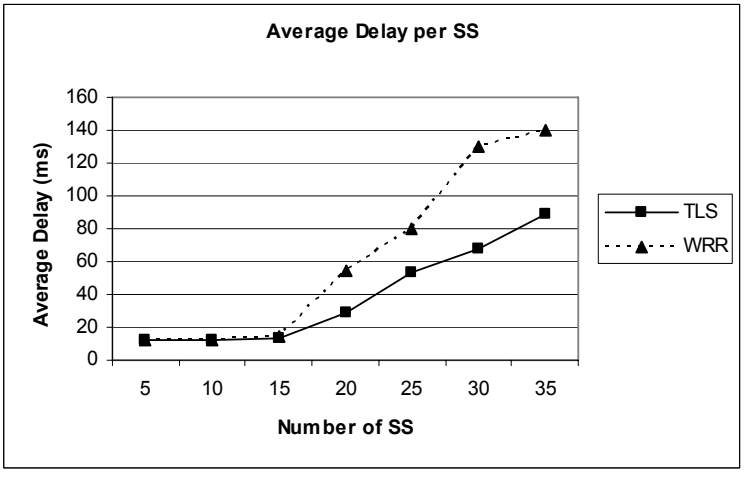

Fig. 4. Average Delay

time service applications, however, for the WRR scheduler algorithm, the scheduler delay is more than 140ms, and it cannot satisfy the demand of real time service applications.

4.3.3 Packet Loss

Fig. 5 displays packet loss, which is the percentage of packets dropped in transit. In an error condition, data packets appear to be transmitted correctly at one end of a connection, but never arrive at the other. Packet loss can be caused by a number of factors, including signal degradation, channel congestion, or corrupted packets rejected in-transit. The main observable effect of packet loss is lower data throughput and poor performance for the real time services application. The packet loss increases with increasing numbers of SSs, TLS algorithms indicate lower average delay than the RR Algorithm. Because the real-time-services application has higher priority

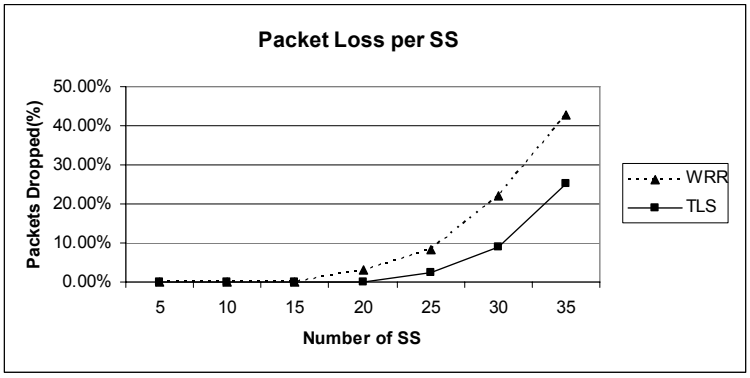

Fig. 5. Packet Loss

than non-real-time-services application, we schedule the real-time-services application first, so the packet loss of the real-time-services application is lower.

5. CONCLUSION

In conclusion, compared with the Round Robin algorithm and Weighted Round Robin algorithms, the QoS Priority and Fairness scheduling scheme for downlink traffic guarantees the delay requirement of UGS and ertps and rtps service flows and maximizes the throughput of BE service flows in the downlink. A simulation study was used to compare the performance of the QoS Priority and Fairness scheduling scheme and the Round Robin and weighted round robin algorithms. The simulations verified that the QoS Priority and Fairness scheduling scheme has the highest throughput and the minimum delay for high QoS classes. For future work, we will compare the TLS scheme with other scheduling algorithms, such as deficit round robin (DRR), earliest deadline first (EDF) and weighted modified deficit round robin (MDRR). Moreover, we will research admission control mechanisms, which decide whether the new connection is established or not. This mechanism guarantees the QoS of current connection, and also ensures the QoS of the newly accepted connection. When a new connection is accepted, there should be enough available bandwidth to accept the new connection. If the connection is accepted, the system will guarantee the QoS parameter of the new connection, such as delay and jitter. The QoS of the current connection will not be influenced by the new connection.

REFERENCE

[1] IEEE P802.16Rev2/D2, DRAFT Standard for Local and metropolitanarea networks, "*Part 16: Air Interface for Broadband Wireless Access Systems*", pp. 2094, December 2007.

[2] WiMAX Forum, "WiMAX System Evaluation Methodology V2.0", pp. 230, December 2007.

[3] So-In, C., Jain C. R., Al-Tamimi, A., "Scheduling in IEEE 802.16e Mobile WiMAX Networks:Key Issues and a Survey", IEEE Journal on Selected Areas in Communications (JSAC), Vol. 27, No. 2, pp. 156-171, 2009.

[4] Martello, S., Vigo, D., "*Exact solution of the two-dimensional finite bin packing problem*", *Manage. Sci.,* Vol. 44, No. 3, pp. 388-399, 1998.

[5] Wäscher Wäscher, G., Haußner, H., Schumann, H., "*An improved typology of cutting and packing*

problems", Eur. J. Operat. Res., Vol. 183, No. 3, pp. 1109-1130, 2007.

[6] Korf., R. E., "Optimal rectangle packing: New results. In Shlomo Zilberstein, Jana Koehler, and Sven Koenig, editors", ICAPS, pp. 142–149, AAAI, 2004.

[7] Moffitt, M. D., Pollack, M. E., *"Optimal rectangle packing: A meta-CSP approach. In Derek Long, Stephen F. Smith, Daniel Borrajo, and Lee McCluskey, editors", ICAPS,* pp. 93–102, AAAI, 2006.

[8] Garey, M-R., and Johnson, D-S., *"Computers and Intractability: A Guide to the Theory of NPCompleteness", W. H. Freeman and co.,* pp. 340, January 1979.

[9] So-In, C., Jain, R., Al-Tamim,i A., *"Generalized Weighted Fairness and its support in Deficit Round Robin with Fragmentation in IEEE 802.16 WiMAX," The 2nd International Conference on computer and Automation Engineering (ICCAE)",* Singapore, Vol. 1, pp. 784-788, 2010.

[10] Katevenis, M., Sidiropoulos, S., Courcoubetis, C., *"Weighted Round-Robin Cell Multiplexing in a General-Purpose ATM Switch Chip," IEEE J. Selected Areas in Comm.,* Vol. 9, No. 8, pp. 1265-1279, 1991.

[11] Li, B., Qin,Y., Low, C. P., Gwee, C. L., *"A Survey on Mobile WiMAX", IEEE In Communications Magazine,* Vol. 45, No. 12, pp. 70-75, 2007.

[12] http://www.opnet.com/solutions/network_rd/modeler_wireless.html

Wei Nie

He received an M.S. degree from University of Electronic Science and Technology of China (UESTC) in 2004. Now he is a doctoral candidate in University of Electronic Science and Technology of China (UESTC). His current research interests include Ad Hoc Network, Sensor Network and WiMAX.

Houjun Wang

Dr. Houjun Wang is a professor of Electronic Science and Technology of China. In recent years, the main research interests include industrial applications of wearable computing, portable testing technology for field application, the next test computing system architecture, etc.

Jong Hyuk Park

Dr. James J. (Jong Hyuk) Park received his Ph.D. degree in Graduate School of Information Security from Korea University, Korea. From December, 2002 to July, 2007.He is now a professor at the Department of Computer Science and Engineering, Seoul National University of Science and Technology, Korea.