

A Latency and Modulation Aware Bandwidth Allocation Algorithm for WiMAX Base Stations

Yi-Neng Lin, Che-Wen Wu, Ying-Dar Lin
National Chiao-Tung University
Hsinchu, Taiwan

Yuan-Cheng Lai
National Taiwan University of Science and Technology,
Taipei, Taiwan

Abstract—The mobile WiMAX systems based on IEEE 802.16e-2005 provide high data rate for the mobile wireless network. However, the link quality is frequently unstable owing to the long-distance and air interference and therefore impacts real-time applications. Thus, a bandwidth allocation algorithm is required to be modulation-aware, while further satisfying the latency guarantee, service differentiation and fairness. This work proposes the Highest Urgency First (HUF) algorithm to conquer the above challenges by taking into consideration the adaptive modulation and coding scheme (MCS) and the urgency of requests. Downlink and uplink sub-frames are determined by reserving the bandwidth for the most urgent requests and proportionating the remaining bandwidth for others. Then, independently in the downlink and uplink, the HUF allocates bandwidth to every mobile station according to a pre-calculated U -factor which considers urgency, priority and fairness. Simulation results prove the HUF is modulation-aware and achieves the above three objectives, notably the zero violation rate within system capacity as well as the throughput paralleling to the best of the existing approaches.

I. INTRODUCTION

IEEE 802.16 [1], known as WiMAX, is an emerging next-generation mobile wireless technology standardized based on the cable network protocol, DOCSIS [2] from which it inherits some features such as the point-to-multipoint system architecture, Quality of Service (QoS) service classes. Different from its predecessor, WiMAX transmits data over the air interface rather than over the cable, so that mobility further specified in the 802.16e-2005 [3], can be supported. The widely used Wi-Fi [4] is point-to-multipoint and also supports mobility. However, WiMAX has separate downlink (DL) and uplink (UL) channels to utilize the bandwidth efficiently and to alleviate the lengthy contention delay. To accomplish these, WiMAX has a control center named base station (BS) for managing the DL/UL transmissions and allocating bandwidth for mobile stations (MSs), rather than arbitrary contentions adopted in Wi-Fi.

With the ever-growing bandwidth demand of time-sensitive multimedia applications, the bandwidth in wireless environment becomes relatively scarce. Though service classes and parameters such as minimum reserved rate, maximum sustained rate and maximum latency, have been defined in the standard for service differentiation, an appropriate bandwidth allocation algorithm is required in BS to achieve satisfactory quality along with the following considerations. First, the *Grant Per Subscribe Station* (GPSS) scheme which is mandatory in the

standard and more flexible than the *Grant Per Connection* (GPC) in the DOCSIS. In GPSS the BS grants requested bandwidth per MS rather than per connection so that the MS can respond to connections of different QoS requirements. Second, the modulation types and coding schemes (MCS) of BS to every MS shall be adaptive to the distance and air interference. The MCS decides the transmission data rate and the translation from bytes to physical slots. Third, among other QoS requirements, the *maximum latency* is most critical to the quality of time-sensitive multi-media applications and thus should be properly satisfied.

A number of designs have been proposed to deal with the above-mentioned considerations. The MLWDF [5] is throughput-optimal and using the head-of-line waiting time of packet as scheduling metric for real time traffic, but the QoS service classes are not involved. The UPS [6] and DFPQ [7] employ service classes to meet differentiation and fairness, while the TPP [8] further uses the dynamic adjustment of the downlink (DL) and uplink (UL) to maximize the bandwidth utilization. However, they do not concern the physical-layer characteristics such as MCS. In [9], the authors cover this and Strict Priority is applied, though latency is ignored and starvation could occur easily for the low-level service classes. Although those solutions are innovative, an integrated algorithm is demanded.

In this work, a bandwidth allocation algorithm, *Highest Urgency First* (HUF), is proposed to tackle those challenges with the physical-layer being OFDMA-TDD, which is the most prevalent physical-layer technology for the WiMAX systems. The algorithm consists of four steps: (1) translating the data bytes of requests to slots reflecting the MCS of every MS, and calculating the number of frames to satisfy the maximum latency for every request of the service flows; (2) pre-calculating the number of slots required by DL/UL requests which must be transmitted in these scheduled frame, and then deciding the portion of DL/UL sub-frame; (3) allocating the slots for every flow using U -factor, which considers the urgency of every bandwidth request, and (4) allocating the slots of every queue to MSs.

The rest of this work is organized as follows. Section II briefs the 802.16 PHY and MAC features and reviews related studies to justify our problems. Section III describes the detailed procedures of the proposed algorithm. Section IV presents the simulation environments and evaluation results. Finally, section V concludes this work with some future directions.

II. BACKGROUND

A. PHY Layer Features

Orthogonal Frequency Division Multiple Access (OFDMA), extended based on the Orthogonal Frequency Division Multiplexing (OFDM), further supports multiple accesses. Resources are available in OFDMA in the time domain in terms of symbols and in the frequency domain in terms of sub-carriers which are grouped into sub-channels. The minimum frequency-time resource unit is one slot which is equal to 48 data sub-carriers and the number of symbols used in a slot is called slot duration, which contains two symbols for DL while three symbols for UL in the mandatory PUSC mode. The mobile WiMAX adopts OFDMA for improving multi-path performance in non-line-of-sight environment. 802.16 PHY supports Time Division Duplex (TDD), Frequency Division Duplex (FDD), and Half-Duplex FDD modes. However, the TDD is preferred in WiMAX since it only needs one channel, enabling the adjustment of unbalanced DL/UL loads, while the FDD needs two channels.

An OFDMA-TDD frame is composed of (1) preamble for synchronization, (2) DL-MAP and UL-MAP for control and element information describing bursts for all MSs, and (3) the DL/UL data bursts carrying data for MSs. The amount of data carried in a slot varies with different adaptive modulations and coding schemes (MCS) which decides the transmission data rate according to the link quality between the BS and MSs.

B. MAC Layer with QoS

Five uplink service classes, the Unsolicited Grant Service (UGS), Real-time Polling Service (rtPS), Non-real-time Polling Service (nrtPS), Best Effort (BE), and the replenished Extended Real-time Polling Service (ertPS) are supported in the 802.16e-2005. Table I summarizes the characteristics of those service classes.

TABLE I. Service classes and the corresponding QoS parameters.

Feature	UGS	ertPS	rtPS	nrtPS	BE
Request Size	Fixed	Fixed but changeable	Variable	Variable	Variable
Unicast Polling	N	N	Y	Y	N
Contention	N	Y	N	Y	Y
QoS Parameters	Min. rate	N	Y	Y	N
	Max. rate	Y	Y	Y	Y
	Latency	Y	Y	Y	N
	Priority	N	Y	Y	Y
Application	VoIP without silence suppression, T1/E1	Video, VoIP with silence suppression	Video, VoIP with silence suppression	FTP, Web browsing	E-mail, message-based services

The scheduling flows within BS and MS are elaborated as follows. While the DL scheduler in a BS simply distributes DL data to MSs, the UL scheduler needs to reserve grants for MSs for the UGS and ertPS flows as well as for the UL bandwidth requests of rtPS, nrtPS and BE flows submitted through polling or contention. Notably the QoS parameters are involved in the meantime. The scheduling results are then passed to the frame builder, in which the DL-MAP/UL-MAP is generated. The

DL-MAP/UL-MAP portrays the DL/UL sub-frame information to notify the PHY layer when to send/receive data bursts. As for the MS side, the scheduler schedules the UL data based on the number of granted slots documented in the UL-MAP. Obviously, the bandwidth allocation algorithm exercised by the BS's scheduler is critical and must be designed carefully in order to optimize the system performance.

C. Related Works

A number of works concerning the bandwidth allocation over IEEE 802.16 can be found. Andrews and Kumaran [5] propose the MLWDF to maximize the channel capacity for multiple MSs performing real-time applications to support QoS. It uses the head-of-line packet's waiting time or the total queue length as the scheduling metric for throughput optimality and satisfaction with delay requirement. Wongthavarawat and Ganz [6] propose the *Uplink Packet Scheduling* (UPS) for service differentiation. It exploits the Strict Priority to select the target class to be scheduled, in which each service class adopts a certain scheduling algorithm for its own queues. However, this scheme only concerns the uplink and hence the overall bandwidth is suffered and low priority classes tend to suffer from starvation. The *Deficit Fair Priority Queue* (DFPQ) [7] revises the UPS by replacing the Strict Priority with the use of maximum sustained rate as the deficit counter for the transmission quantum of every service class, and therefore can dynamically adjust the DL and UL proportion according to the counters. Nevertheless, this scheme is suitable only for the GPC mode and setting an appropriate maximum sustained rate is not trivial. The *Two Phase Proportionating* (TPP) [8] introduces a simple approach to dynamically proportionate the DL and UL sub-frames and considers the minimum reserved rate, maximum sustained rate and the requested bandwidth of service classes in terms of the *A-Factor* to grant the bandwidth for MSs proportionally. However, it could lead to inappropriate grants owing to the proportion. All above schemes do not consider the MCS which affects the transmission data rate and the service quality. Sanyenko's approach [9] involves the MCS, but does not provide the latency guarantees.

D. Problem Statement

To integrate all features in WiMAX PHY and QoS service classes and solve the above-mentioned problems, a well-designed algorithm is demanded to satisfy the following metrics. First, it must be aware of the adaptive MCS in PHY and translate the bandwidth of request to appropriate number of slots to meet the bandwidth demand and grant for every MS. Second, service classes must be satisfied for the requirements of QoS parameters such as minimum reserved rate, priority and maximum latency. The maximum latency guarantee is most important for the real time application in rtPS. Third, for fairness, the allocation algorithm should serve the service classes fairly to avoid the starvation of low priority service classes. The problem statement leads to design a modulation,

latency and priority –aware dynamic downlink and uplink bandwidth allocation in a WiMAX BS.

III. HIGHEST URGENCY FIRST

This section elaborates the concept and procedures of the proposed *Highest Urgency First* (HUF) algorithm. The HUF uses the *Urgency* parameter to schedule all requests considering latency guarantee and fairness, and divides the allocation procedure into two phases. The first phase determines the bandwidth of DL/UL sub-frame while the second phase allocates bandwidth for data/bandwidth requests from MSs. Each phase manipulates different metrics to achieve high throughput, latency guarantee and fairness. The components and operations of the HUF algorithm are illustrated in Fig. 1.

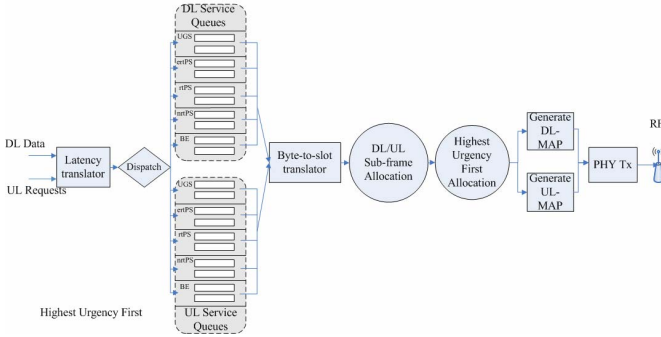


Fig. 1. Procedure of the Highest Urgency First (HUF)

A. Detailed Procedures of the Algorithm

With regard to data/request translation and deadline determination, in the uplink, a service flow in MSs expedites a bandwidth request to BS whenever necessary, while in the downlink data are en-queued, scheduled and finally sent down to MSs. The transmission unit in WiMAX is slot whose capacity depends on the current MCS. Therefore, when a new frame starts, according to the MCS the required data/request size is firstly translated into number of slots as

$$\#_of_slots = \frac{BQS}{bytes_per_slot} \quad (1)$$

where the BQS denotes the data size of the data/request, and $bytes_per_slot$ represents the capacity of a slot. Since a slot contains 48 data sub-carriers in Mobile WiMAX PHY [10] and the MCS decides number of bits carried in a sub-carrier, we can thus have

$$bytes_per_slot = \frac{48 * Mod_bits * Coding_rate}{8} \quad (2)$$

Regarding the service classes such as UGS, ertPS and rtPS, the maximum latency parameter is expected to satisfy for guarantee the quality of real-time applications. Thus, in the algorithm the deadline is defined as

$$deadline = \left\lfloor \frac{ML}{FD} \right\rfloor \quad (3)$$

where ML means the maximum latency of the service flow and FD represents the frame duration. If the maximum latency is not set in the service flow, the deadline of the requests belonging to

that flow is set to -1. Otherwise, the corresponding deadline is calculated upon the arrival of a data/request, and then decreased by one after a frame duration. A deadline equaling to zero indicates the violation of the maximum latency requirement.

In the first phase for DL/UL Sub-Frame Allocation, in order to fill up the frame to achieve high throughput while considering the latency requirement for the service flows, HUF uses the urgent data/requests with deadline equaling to one and non-urgent data/requests which except the urgent ones as the metrics to decide the DL/UL sub-frame size. Besides, the minimum reserved rate is a necessary requirement for every service flow. Therefore, it is also taken to consideration. Detailed procedure to decide the DL/UL ratio is as follows:

- 1) For DL and UL, respectively, sum up the number of data/requests slots whose deadline equals to one in all queues so as to reserve bandwidth for those that must be served in this frame.
- 2) For DL and UL, respectively, sum up the amount of slots translated from the minimum reserved rate of every service flow. Exclude those that have been considered in i.
- 3) Sum up the number of reserved slots calculated from i and ii. Divide them by the number of DL/UL sub-channel in a slot duration to obtain the amount of symbols to be reserved. Notably in PUSC mode a slot duration spans two symbols in DL yet three in UL.
- 4) The amount of remaining symbols is thus calculated by subtracting the number of reserved symbols from the total number of symbols in a frame. Proportionate the remaining symbols for the DL and UL according to their amount of bandwidth requested by data/requests having deadlines larger than one. Letting DR and UR represent the above requested bandwidth for DL and UL, respectively, the proportion can be derived as

$$\frac{UR}{DR} = \frac{(S_{rem} - (SD_{DL} \times x)) / SD_{UL}}{x} = \frac{S_{rem} - (SD_{DL} \times x)}{SD_{UL} \times x} \quad (4)$$

where S_{rem} indicates the number of remaining symbols and SD_{DL} and SD_{UL} means the number of symbols in a DL and UL slot duration, respectively. x which is the number of slot durations DL obtains can be found after solving the equation, in which $\frac{S_{rem} - (SD_{DL} \times x)}{SD_{UL}}$ represents the amount of slot durations distributed to the UL.

In short, HUF reserves symbols for the data/requests which must be served in this frame, then proportionate the remaining symbols by the non-urgent data/requests to decide the DL/UL sub-frame size.

In the second phase for Highest Urgency First Allocation, After the DL and UL sub-frame sizes are determined in first phase, the HUF scheduler starts to allocate independently the bandwidth of DL/UL sub-frame to MSs. The essence of HUF is to ensure the requirements of maximum latency and priority

among all service flows, and allocate the bandwidth to MSs fairly. Hence, HUF allocates the bandwidth in the precedence based on that requested slots whose deadline is one and satisfying the minimum reserved rate of every flow. Then, when there is bandwidth left in a sub-frame, HUF defines the *U-factor* to select the other data/requests to be served. The allocation procedure in the uplink is portrayed as follows:

- 1) For DL and UL, respectively, sum up the number of data/requests slots whose deadline equals to one in all queues so as to reserve bandwidth for those that must be served in this frame.
- 2) Calculate the *average-U-factor* for every service flow. Flows are subsequently served, by dispatching the head-of-line request only, in decreasing order of *average-U-factor*. The *average-U-factor* of a service flow can be derived as

$$\text{average-U-factor} = \frac{\sum_{i=1}^n U\text{-factor}_i}{n}, \quad \text{where} \quad (5)$$

$$U\text{-factor}_i = \frac{N_i \times (P+1)}{D_i} \quad (6)$$

indicates the Urgency of the i th request in the flow and n represents number of requests. As shown in Eq. 6, the $U\text{-factor}_i$ comprises three metrics, namely D_i , P and N_i . D_i means the deadline of the i th bandwidth request. For flows not having a deadline, the HUF automatically associates them with a value which is the maximum deadline among all UL requests. P stands for the flow priority, which is defined in the 802.16 standard and ranges from zero (lowest) to seven (highest). N_i is the number of slots translated from the requested size. Once the head-of-line requests of all queues are dispatched, the HUF performs step ii, namely recalculating the *average-U-factors* and so forth, repeatedly until the UL sub-frame is fulfilled.

The downlink is treated similarly the uplink. With regard to grant bandwidth for each MS, after allocating bandwidth to requests of each queue, the HUF scheduler further distributes the bandwidth to every MS by totaling up the allocated bandwidth of the service queues of the same MS. Based on the grants, the scheduler generates the corresponding DL and UL MAPs which are sent every frame to notify the MSs when to transmit/receive data. Finally the HUF updates the deadline of every request by $Deadline = Deadline - 1$.

IV. EVALUATION RESULTS

The evaluation scenarios cover the MCS awareness, latency guarantee, and fairness in service classes. Furthermore, only rtPS and BE are involved in the following evaluation because the UGS as well as the ertPS are granted with fixed bandwidth, while the nrtPS differs from the BE merely in the priority.

A. Simulation Environment

The simulation topology is depicted in the Fig. 2. A number of MSs and a BS are connected via a gateway to a video conference endpoint and an FTP server.

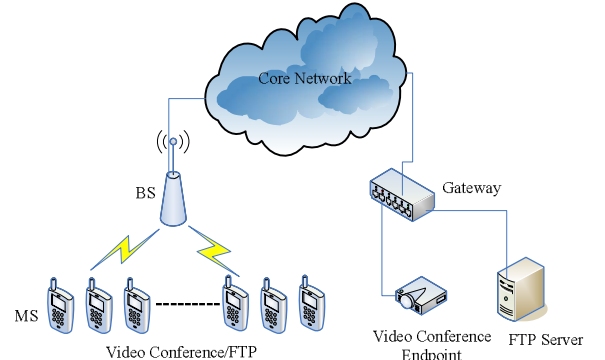


Fig. 2. Simulation topology

The video conference application used in the simulation has variable packet size and is constrained with the latency requirement for maintaining the quality of the rtPS and FTP for the BE. The WiMAX system profile [10] and application parameters are summarized in Table II.

TABLE II
(a) System Profile and (b) Application Parameters in the Simulation

System Parameter	DL	UL
System Bandwidth	1.5 MHz	
FFT Size	1024	
Frame Duration	5 ms	
Useful Symbol Time ($T_b = 1/f$)	60 μ s	
Guard Time ($T_g = T_b/g$)	7 μ s	
OFDMA Symbol Duration ($T_s = T_b + T_g$)	67 μ s	
Sub-channels	10	12
Slots of per sub-channel	1	1
Number of Symbol Duration of per slot	2	3

(b)

Application	Parameter
Video Conference	Packet Size:
	- Lognormal Distribution
	- Average: 4.9bytes
	- Standard deviation: 0.75 bytes [12]
	Packet inter-arrival time: 30 frames/sec
FTP	Requested file size: 200Kbytes
	Inter-request Time: 30 sec

B. Modulation-aware Allocation

Whenever the MCS is changed due to interferences, for

consistent video conferencing quality the data rate of MSs must be sustained by granting each of them adapted number of slots. Table III depicts the modulation awareness of the HUF, in which two MSs whose MCSs change along with time, are involved. From Fig. 3 we observe that though the modulation alters, the throughput is still kept the same. This is because more slots are granted as the capacity of a slot shrinks due to an un-scalable MCS. Similar behavior occurs otherwise.

TABLE III The effect of the changing MCS

Modulation	BPSK		QPSK		16QAM		64QAM	
Coding scheme	1/2	1/2	3/4	1/2	3/4	1/2	2/3	3/4
Bytes/slot	3	6	9	12	18	18	24	27
MS1	0~10	10~15	15~20	20~25	25~30	30~35	35~40	40~50
MS2	40~50	35~40	30~35	25~30	20~25	15~20	10~15	0~10

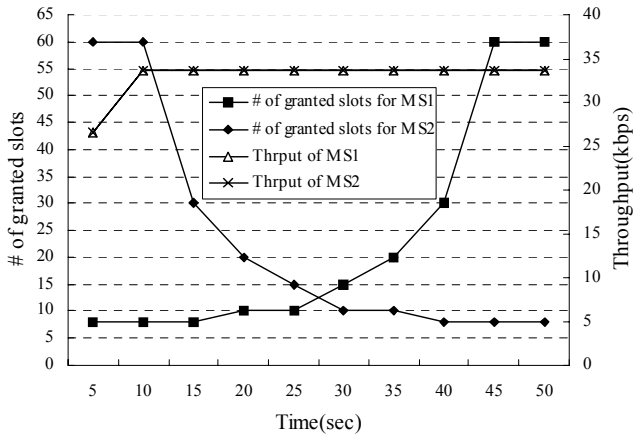


Fig. 3. Modulation-aware allocation: the throughput is kept whenever the MCS is changed.

C. Latency Guarantee with Different Requirements

We compare the proposed algorithm with the MLWDF which is throughput-optimal and considers the waiting time of head-of-line packet to keep the latency requirement, and with the DFPQ which uses EDF [7] for rtPS to satisfy the requirement. The evaluation scenario uses the video conference application referencing Table II(b) based on two sets of QoS parameters used in rtPS presented in Table IV. Among the parameters only one is configured differently, namely the maximum latency requirements which is 50ms and 150ms, respectively. The load of the link is increased by simultaneously increasing the input of two traffic flows.

TABLE IV. QoS parameters of the two types of traffic flows

QoS Parameter	TYPE I	TYPE II
Service Class	rtPS	rtPS
Minimum Reserved Rate (bps)	2400	2400
Maximum Sustain Rate (bps)	1000000	1000000
Maximum Latency (ms)	50	150
Polling Time (ms)	20	20

The criteria of the evaluation are throughput, average latency of packets and violation rate. The throughput and average latency are the general criteria to evaluate the performance of a

bandwidth allocation algorithm. Besides, the evaluation scenario focuses on the satisfaction with different latency requirements, and thus takes the violation rate into account. Latency guarantee means the violation rate is zero regardless of the requirements. Figure 4 discusses the throughput as well as the latency of three algorithms.

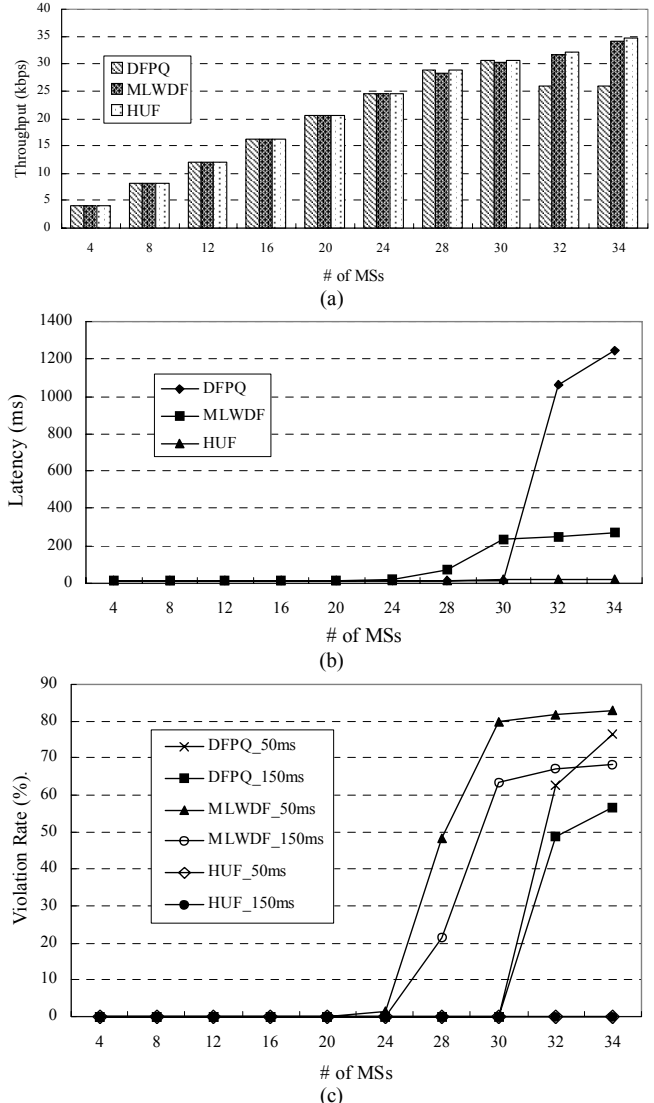


Fig. 4. (a) Throughput, (b) average latency and (c) violation rate of three different algorithms

Fig. 4(c) further examines the violation of the three algorithms in latency. Even when the number of MSs is up to 34, the HUF has no violation for the maximum latency being 50ms and 150ms. Nevertheless, the violation rate of MLWDF grows drastically when 28 MSs are involved and is close to 70% and 80%, respectively for maximum latency requirement being 150ms and 50ms when 34 MSs are present. The DFPQ has a violation rate of 58% for 50ms and 78% for 150ms for 34 MSs resulted from the degraded throughput.

C. Fairness

Herein we evaluate the fairness of the proposed algorithm HUF with DFPQ, TPP, and UPS. In the evaluation two sets of MSs are involved, in which one performs rtPS-based video conferencing and the other uploads files via BE-based FTP. The application profiles are shown in Table II(b) while the parameters of service classes are presented in Table V.

TABLE V
The parameters of rtPS and BE

QoS Parameter	TYPE I	TYPE II
Service Class	rtPS	BE
Minimum Reserved Rate (bps)	2400	2400
Maximum Sustain Rate (bps)	1000000	1000000
Maximum Latency (ms)	50	N/A
Polling Time (ms)	20	N/A

The fairness between rtPS and BE can be formulated as

$$Fairness_{r-b} = \left| \frac{Th_{rtPS}}{S_{rtPS}} - \frac{Th_{BE}}{S_{BE}} \right| \quad [7] \quad (7)$$

where S_{rtPS} and Th_{rtPS} are the requested bandwidth and the corresponding throughput of rtPS, yet S_{BE} and Th_{BE} are those of BE. Figure 5(a) shows that TPP and HUF are fairer than DFPQ and UPS. That is because the UPS uses Strict Priority to allocate bandwidth to all service classes in which BE tends to get starved as the rtPS becomes demanding. In DFPQ, the maximum sustained rate is employed as the *Deficit* counter; however deciding the appropriate maximum sustained rate for all service classes is not trivial. Thus, if the maximum sustained rate is not configured properly, the fairness is degraded. Fig. 5(b) further explains the results. As shown in the figure, all approaches allocate fairly, namely 17% for rtPS and 83% for BE, when 4 MSs are employed. However, UPS and DFPQ start to distribute excessive number of slots to rtPS for 8 MSs because of higher priority, resulting in the starvation of BE. Contrastively, HUF is quite fair even when 16MSs are involved.

V. CONCLUSION

In this paper, the *Highest Urgency First* (HUF) is proposed. It translates the requested size to number of slots according to the current MCS when a frame starts, and then allocates the bandwidth according to the *Urgency* of the data/request. Simulation result indicates that the quality is retained as the MCS adapts owing to the link quality. HUF outperforms the DFPQ by 25% in throughput when overloaded, and incurs no latency violation when the load is within system capacity. Finally, we compare the fairness of UPS, DFPQ, TPP and HUF and observe fairness between rtPS and BE in HUF which, unlike the TPP, avoids inappropriate grant for rtPS.

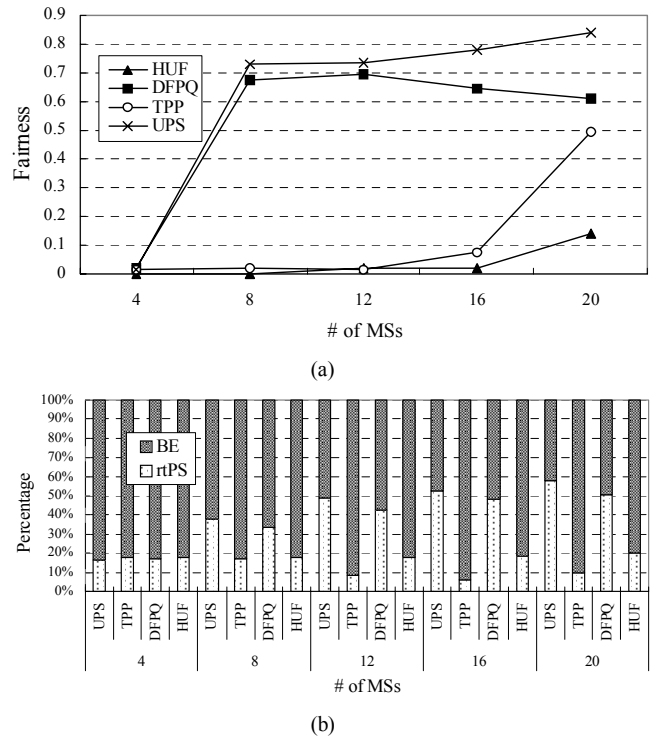


Fig. 5. a) Fairness and b) granted slots for rtPS and BE of four algorithms

REFERENCES

- [1] IEEE 802.16 Working Group, "Air interface for fixed broadband wireless access systems," Jun 2004.
- [2] Cable Television Laboratories Inc., "Data-Over-Cable Service Interface Specifications - Radio Frequency Interface Specification v1.1," July 1999.
- [3] IEEE 802.16 Working Group, "Air Interface for Fixed and Mobile Broadband Wireless Access Systems - Amendment for Physical and Medium Access Control Layers for Combined Fixed and Mobile Operation in Licensed Bands," Feb. 2006.
- [4] IEEE 802.11 Working Group, "Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications," Sep. 1999.
- [5] M. Andrews et al., "Providing Quality of Services over a Shared Wireless link," IEEE Communication Magazine, Feb. 2001, pp. 150-154.
- [6] K. Wongthavarawat, A. Ganz, "IEEE 802.16 Based Last Mile Broadband Wireless Military Networks with Quality of Service Support," MILCOM, Oct. 2003.
- [7] J. Chen, W. Jiao, H. Wang, "A Service Flow Management Strategy for IEEE 802.16 Broadband Wireless Access Systems in TDD Mode," ICC, May 2005.
- [8] Y. N. Lin, S. H. Chien, Y. D. Lin, Y. C. Lai, M. Liu, "Dynamic Bandwidth Allocation for 802.16e-2005 MAC," Book Chapter of "Current Technology Developments of WiMax Systems," edited by Maode Ma, to be published by Springer, 2007.
- [9] A. Sayenko, O. Alanen, J. Karhula, T. Hamalainen, "Ensuring the QoS Requirements in 802.16 scheduling," ACM MSWiM '06, Oct. 2006.
- [10] Mobile WiMAX Part I, "A Technical Overview and Performance Evaluation," WiMAX Forum, April 2006.