Computer Communications 32 (2009) 332–342

Contents lists available at ScienceDirect

Computer Communications

journal homepage: www.elsevier.com/locate/comcom

Highest Urgency First (HUF): A latency and modulation aware bandwidth allocation algorithm for WiMAX base stations

Yi-Neng Lin [a,*], Ying-Dar Lin [a], Yuan-Cheng Lai [b], Che-Wen Wu [a]

[a] National Chiao-Tung University, 1001 University Road, Hsinchu, Taiwan 300, ROC
[b] National Taiwan University of Science and Technology, 43, Sec. 4, Keelung Rd., Taipei, 106 Taiwan, ROC

ARTICLE INFO

Article history:
Received 14 December 2007
Received in revised form 26 October 2008
Accepted 2 November 2008
Available online 12 November 2008

Keywords:
WiMAX
Bandwidth allocation
Algorithm
Latency
Modulation

ABSTRACT

The mobile WiMAX systems based on IEEE 802.16e-2005 provide high data rate for mobile wireless networks. However, the link quality is frequently unstable owing to mobility and air interference and therefore impacts the latency requirement of real-time applications. In the WiMAX standard, the modulation/coding scheme and the boundary of uplink/downlink sub-frames could be adjusted subject to channel quality and the traffic volume, respectively. This provides us a chance to design a MAC-layer uplink/downlink bandwidth allocation algorithm that is QoS/PHY-aware.

This work takes into account the adaptive modulation and coding scheme (MCS), uplink and downlink traffic volume, and QoS parameters of all five defined service classes to design a bandwidth allocation algorithm that calculates the slot allocation in two phases. The first phase decides the boundary of uplink and downlink sub-frames by satisfying requests with pending latency violation and proportionating according to traffic volume, while the second phase allocates slots to mobile stations considering urgency, priority and fairness. Simulation results show our algorithm achieves zero latency violation and higher system throughput compared to existing non-QoS/PHY-aware or less-QoS/PHY-aware approaches.

© 2008 Elsevier B.V. All rights reserved.

1. 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 and 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, arbitrary contentions for bandwidth lengthen the delay and degrade efficiency. To tackle this, WiMAX further separates the air interface into downlink (DL) and uplink (UL) channels and adopts a control center named base station (BS) for managing the DL/UL transmissions and allocating bandwidth for mobile stations (MSs[1]).

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 [5]. In GPSS the BS grants bandwidth to a MS rather than to certain connection, so that the MS can respond to connections of different QoS requirements. Second, the modulation types and coding schemes (MCS) which decides the data rate between BS and MS and the translation from bytes to physical slots, shall be adaptive to the varying distance and air interference. Third, among all QoS requirements, the *maximum latency* is most critical to the quality of time-sensitive multimedia applications and thus should be properly satisfied.

A number of designs have been proposed to deal with the above-mentioned considerations. The MLWDF (Modified Largest Weighted Delay First) [6] is throughput-optimal and using the waiting time of head-of-line packet as scheduling metric for real time traffic, but the QoS service classes are not involved. The Uplink Packet Scheduling (UPS) [7] and DFPQ (Deficit Fair Priority Queue) [8] employ service classes to meet differentiation and fairness, while the TPP [9] further uses the dynamic adjustment of the downlink (DL) and uplink (UL) to maximize the bandwidth

* Corresponding author. Tel.: +886 922511430.
E-mail address: ynlin@cs.nctu.edu.tw (Y.-N. Lin).

[1] The terminal station is named subscribe station (SS) in the standard 802.16d-2004 for fixed systems, and mobile station (MS) in the standard 802.16e-2005. Below we use MS to represent the terminal station.

0140-3664/$ - see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.comcom.2008.11.003

utilization. However, they do not concern the physical-layer characteristics such as MCS. In [10], the authors cover this and Strict Priority is applied, though latency is ignored and starvation could occur easily for the low-level service classes even an admission control scheme is installed.

In this work, a bandwidth allocation algorithm, *Highest Urgency First* (HUF), is proposed to tackle those challenges with Orthogonal Frequency Division Multiple Access with Time Division Duplex (OFDMA-TDD). OFDMA-TDD, the most prevalent physical-layer technology for the WiMAX systems, has high capacity owing to the OFDMA technique and flexibility in the mobile environment. 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 subframe; (3) allocating the slots for every flow using the *U-factor*, which indicates the latency, priority and fairness of every bandwidth request, and (4) allocating the slots of flows to the corresponding MSs.

The rest of this work is organized as follows. Section 2 briefs the 802.16 PHY and MAC features and reviews related studies to justify our problems. Section 3 describes the detailed procedures of the proposed algorithm. Section 4 presents the simulation environments and evaluation results. Finally, Section 5 concludes this work with some future directions.

2. Background

Since the WiMAX supports high data rate and long distance in the mobile environment, rather than pure contention among MSs which causes significant re-transmissions, a BS must coordinate the decision of transmissions from/to MSs which involves operations in PHY and MAC. In this section, we sketch the WiMAX PHY features which affect the transmission data rate and therefore the bandwidth allocation, and describe the QoS consideration and scheduling flow in the WiMAX MAC. Some related works investigating the allocation problems are discussed, leading to the statement of the research goals.

2.1. Overview of the WiMAX system

2.1.1. PHY layer features

Orthogonal Frequency Division Multiplexing (OFDM) is a multiplexing technology that sub-divides the bandwidth into multiple frequency sub-carriers and exploits the frequency diversity of the multi-path channel by coding and interleaving the information across the sub-carriers prior to transmission. The OFDMA, extended based on the 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 contains 48 data sub-carriers [11] and a slot duration of two symbols for DL while three symbols for UL in the mandatory PUSC (Partial Usage of Sub-Channels) mode. The 802.16 PHY supports 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. Besides, the design of a transceiver is easier in TDD than in FDD [11].

As shown in Fig. 1, 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 MCS which decides the transmission data rate according to the link quality between the BS and MSs. Table 1 summarizes the number of bytes in a slot for all supported MCSs in WiMAX. As an example, the slot capacity when BPSK and code rate of 1/2 are used is 48(bits) $\times$ 1/2 = 3(bytes) since a sub-carrier under BPSK carries 1 bit.

2.1.2. MAC layer with QoS

Five uplink service classes, the Unsolicited Grant Service (UGS), Real-time Polling Service (rtPS), Non-real-time Polling Service

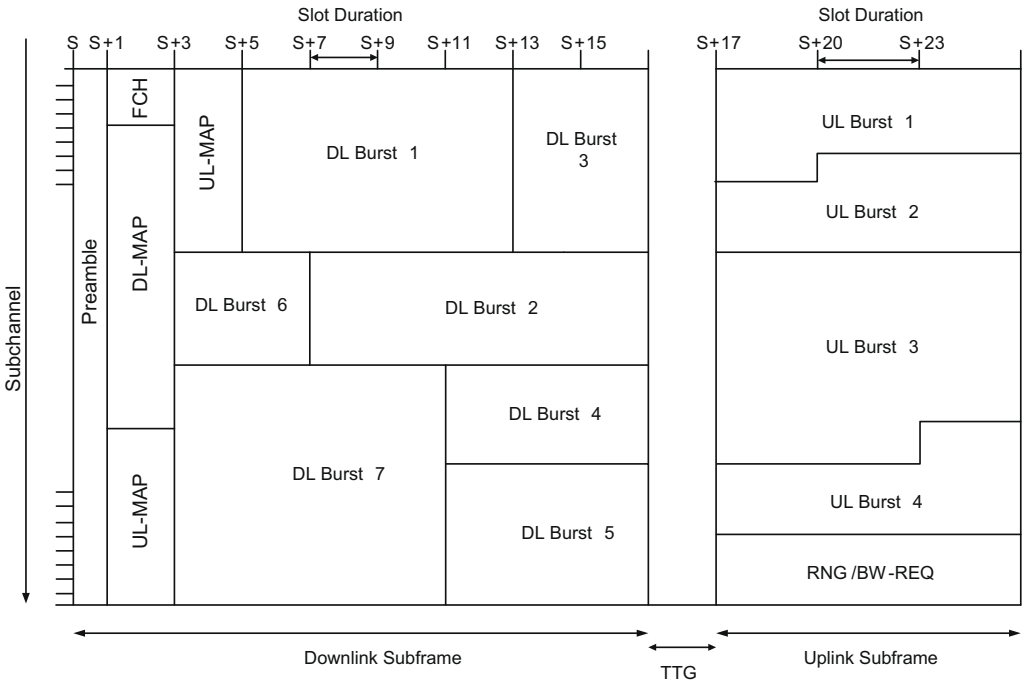

Fig. 1. Structure of a WiMAX OFDMA-TDD frame.

Table 1
Slot sizes of different MCSs in WiMAX.

Modulation	BPSK				QPSK				16QAM				64QAM			
Code rate	1/2	2/3	3/4	5/6	1/2	2/3	3/4	5/6	1/2	2/3	3/4	5/6	1/2	2/3	3/4	5/6
Bytes	3	4	4.5	5	6	8	9	10	12	16	18	20	18	24	27	30

(nrtPS), Best Effort (BE), and the replenished Extended Real-time Polling Service (ertPS) are supported in the 802.16e-2005. A BS reserves bandwidth for UGS flows observing the maximum sustained rate, whereas for rtPS flows it polls the MSs periodically according to the pre-determined time interval and receives bandwidth requests for further allocation. ertPS flows are treated similarly to UGS except that MSs which the flows belong to can further change the reservation size either by contending for chances or using piggyback request field of management packets. nrtPS and BE contend for the transmission opportunities, but nrtPS has extra opportunities to be polled, while BE depends only on contention. Among all service classes except the UGS and ertPS which are provided with sufficient bandwidth, the rtPS must be much concerned since it supports real-time applications having the maximum latency requirement and variable packet sizes. Table 2 summarizes the characteristics of these service classes.

The scheduling flows within BS and MS are shown in Fig. 2 and 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. 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.

2.2. Related work

A number of works concerning the bandwidth allocation over IEEE 802.16 can be found. Andrews and Kumaran [6] 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 [7] 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) [8] 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) [9] 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 [10] involves the MCS, but does not provide the latency guarantees.

2.3. 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 requested bandwidth to appropriate number of slots to meet the bandwidth demand. Second, the QoS requirements of service classes, such as minimum reserved rate, priority and maximum latency, need to be satisfied. Among them the maximum latency guarantee is most important for real time applications belonging to the rtPS class. Third, for fairness, the allocation algorithm should serve the service classes fairly to avoid starvation, given the presence of an admission control scheme. In this article, an operational admission control is assumed, since without it all bandwidth allocation algorithms will inevitably subject to starvation. The problem statement leads to designing a modulation, latency and priority-aware downlink and uplink bandwidth allocation in a WiMAX BS.

3. Highest Urgency First

This section elaborates the concept and procedures of the proposed *Highest Urgency First* (HUF) algorithm. The HUF uses an urgency parameter considering latency, fairness and priority to

Table 2
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	Y	N
Max. rate	Y	Y	Y	Y	Y
Latency	Y	Y	Y	N	N
Priority	N	Y	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

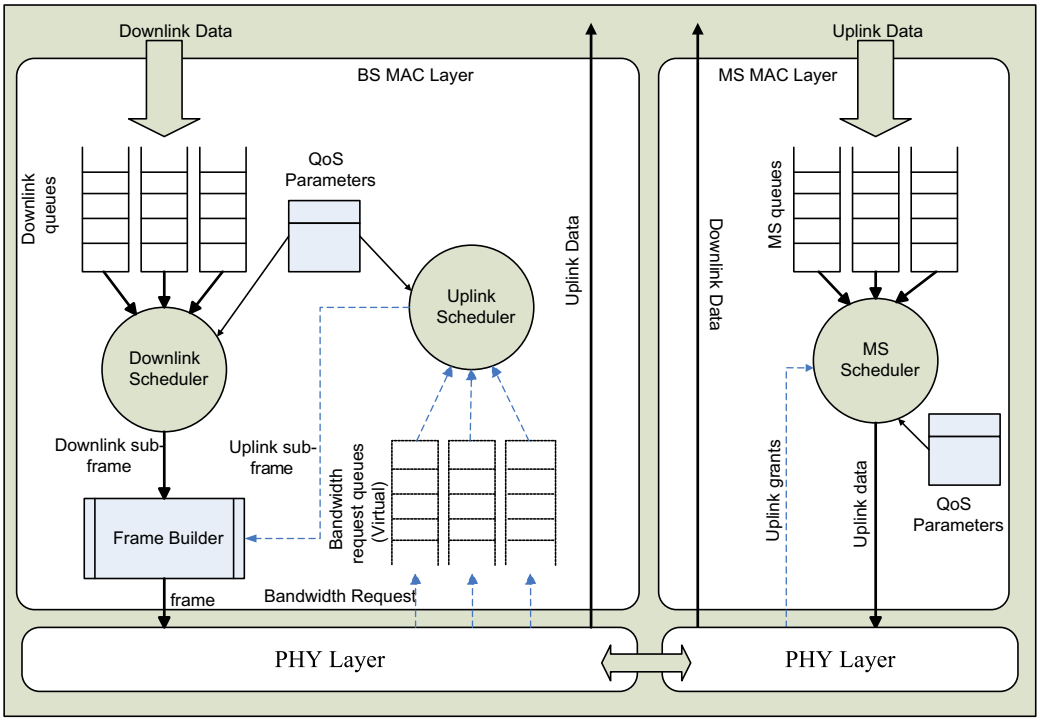

Fig. 2. Scheduling flow and QoS within BS and MSs.

schedule all requests, 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 requests from MSs. Each phase manipulates different metrics to achieve high throughput, latency guarantee and fairness.

3.1. Overview of the algorithm

Our approach aims at dynamically adjusting and filling up the DL/UL sub-frames in TDD mode, while each sub-frame is further allocated to service queues of different QoS requirements such as latency, priority and fairness. Slots in a frame can carry different amount of data owing to the MCS in PHY, and the varying data rate may further affect how bandwidth allocation is performed. Based on the above characteristics, the HUF is proposed to well utilize the bandwidth. An urgency parameter which considers three metrics, namely latency in terms of deadline, fairness in terms of number of requested slots and priority of service flows, is used to decide the servicing order of all data/requests. The deadline represents the number of frame durations left before an uplink request or a downlink packet must be served. A request having a deadline equaling to one must be dispatched in this frame so as to satisfy the latency requirement. The other two contribute to the urgency in terms of the urgency factor, i.e. *U-factor*, in which a higher value indicates a more urgent request. While the priority is trivial as being a metric, the rationale behind the employment of number of requested slots is that, requests demanding large amount of bandwidth shall be allocated as early as possible. They are relatively hard to be scheduled compared to requests of small amount and therefore tend to miss the deadline.

The HUF consists of two phases, first of which decides the size of DL/UL sub-frames based on the minimum reserved rate, the data/requests whose deadline equals to one and other non-urgent demand, while in the second phase DL and UL independently dispatches its own bandwidth to the individual queues of DL and UL according to the minimum reserved rate of every service queue, data/requests in queue whose deadline equals to one, and the *U-factor* of the data/request. Finally, the HUF follows GPSS by granting the accumulated bandwidth of flows to the corresponding MSs. The components and operations of the HUF algorithm are illustrated in Fig. 3 and explained in Section 3.2.

3.2. Detailed procedures of the HUF algorithm

3.2.1. 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 a slot whose capacity depends on the current MCS. Therefore, when a new frame starts, according to the MCS the required size of data/requests is firstly translated into number of slots as

$$\#\_of\_slots = \frac{BQS}{bytes\_per\_slot}, \tag{1}$$

where the *BQS* denotes the requested size, and *bytes_per_slot* represents the capacity of a slot. Since a slot contains 48 data sub-carriers in Mobile WiMAX PHY [11] and the MCS decides the number of bits carried in a sub-carrier, we can thus have

$$bytes\_per\_slot = \frac{48 * Mod\_bits * Coding\_rate}{8}. \tag{2}$$

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

$$deadline = \left\lfloor \frac{ML}{FD} \right\rfloor, \tag{3}$$

where *ML* means the maximum latency for 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

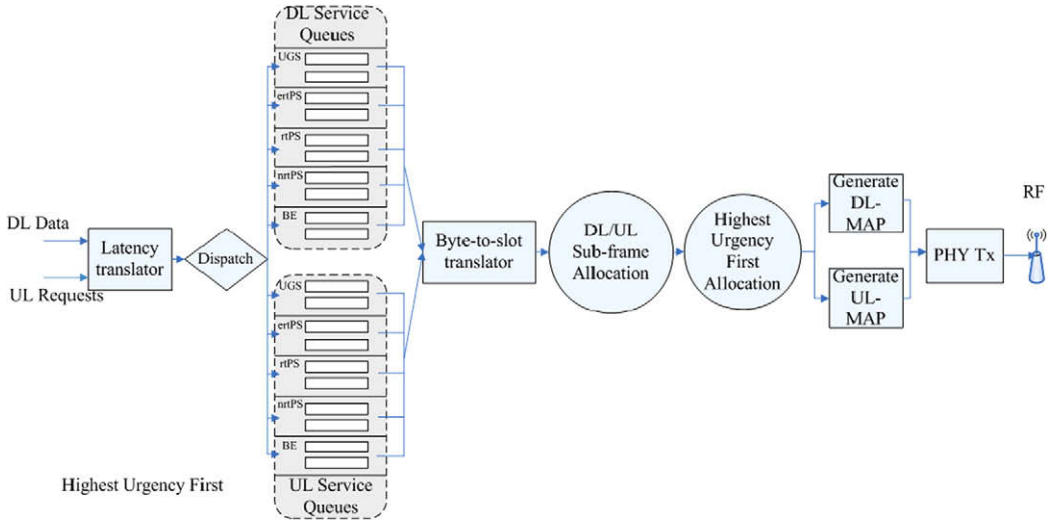

Fig. 3. Procedure of the Highest Urgency First (HUF).

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.

3.2.2. First phase: 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, the HUF uses the urgent data/requests with deadline equaling to one and non-urgent data/requests as the metrics to decide the DL/UL sub-frame size. Besides, the minimum reserved rate is also necessarily considered. Detailed procedure to decide the DL/UL ratio is as follows:

(i) For DL and UL, respectively, sum up the number of data/request slots whose deadline equals to one in all queues so as to reserve bandwidth for those that must be served in this frame.
(ii) 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.
(iii) 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.
(iv) 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} represents 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 $\frac{S_{rem}-(SD_{DL}\times x)}{SD_{UL}}$ specifies the amount of slot durations distributed to the UL.

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

3.2.3. Second phase: urgency-based bandwidth 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:

(i) For each service flow, allocate bandwidth firstly to requests whose deadline equal to one and then to others until the minimum reserved rate is complemented.
(ii) Calculate the *average-U-factor* for every service flow.
(iii) Identify the flow with the highest *average-U-factor* and serve its head-of-line request. Recalculate the *average-U-factor* and repeat step iii until.

The *average-U-factor* of a service flow can be derived as

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

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

indicates the urgency of the ith 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 ith 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 to the uplink.

3.2.4. Grant bandwidth to MSs

After allocating bandwidth to data/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 of the transmission and receiving opportunities. Finally the HUF updates the deadline of every data/request by $Deadline = Deadline - 1$.

Table 3
(a) DL/UL requested slots and (b) system profile in the example.

Direction	Type	Number of requested slots
(a)		
UL (sum of all UL queues)	*deadline* = 1	64
	deadline = 2	40
	deadline = 5	30
	deadline = 7	25
	Min. Rev. Rate	34 + 30 + 20 − 64 = 20
DL (sum of all DL queues)	*deadline* = 1	90
	deadline = 3	50
	deadline = 4	50
	deadline = 6	20
	Min. Rev. Rate	20

PHY parameter	Value
(b)	
Number of symbols in a frame	72
Number of UL sub-channels	12
Number of DL sub-channels	10
UL slot duration (symbols)	3
DL slot duration (symbols)	2

3.2.5. Example

This section elaborates an example of the HUF, in which the parameters and system profile frequently adopted [11] are shown in Table 3. It is assumed that both DL and UL have four queues and the minimum reserved rates are 34, 30, and 20 slots for Queue 1, 2, and 3, respectively. In the first phase, HUF decides the DL/UL sub-frame sizes. According to the requests whose deadline equals to one and the aggregated number of slots for the minimum reserved rate of all DL/UL flows, $((90 + 20)/10) \times 2 = 22$ and $((64 + 20)/12) \times 3 = 21$ symbols are reserved for DL and UL, respectively, with $72 - 21 - 22 = 29$ symbols remained. The HUF then proportionates the remaining symbols for DL and UL by solving Eq. (4) where DR is $50 + 50 + 20 - 20 = 100$ and UR is $40 + 30 + 25 - 20 = 75$. So, DL obtains additional $x = \frac{100\times29}{75\times3+100\times2} \cong 7$ slot durations equaling to $2 \times 7 = 14$ symbols and UL obtains additional $\frac{29-2\times7}{3} = 5$ slot durations equaling to $3 \times 5 = 15$ symbols. Finally, the sizes of DL and UL sub-frames are $22 + 14 = 36$ and $21 + 15 = 36$ symbols, respectively.

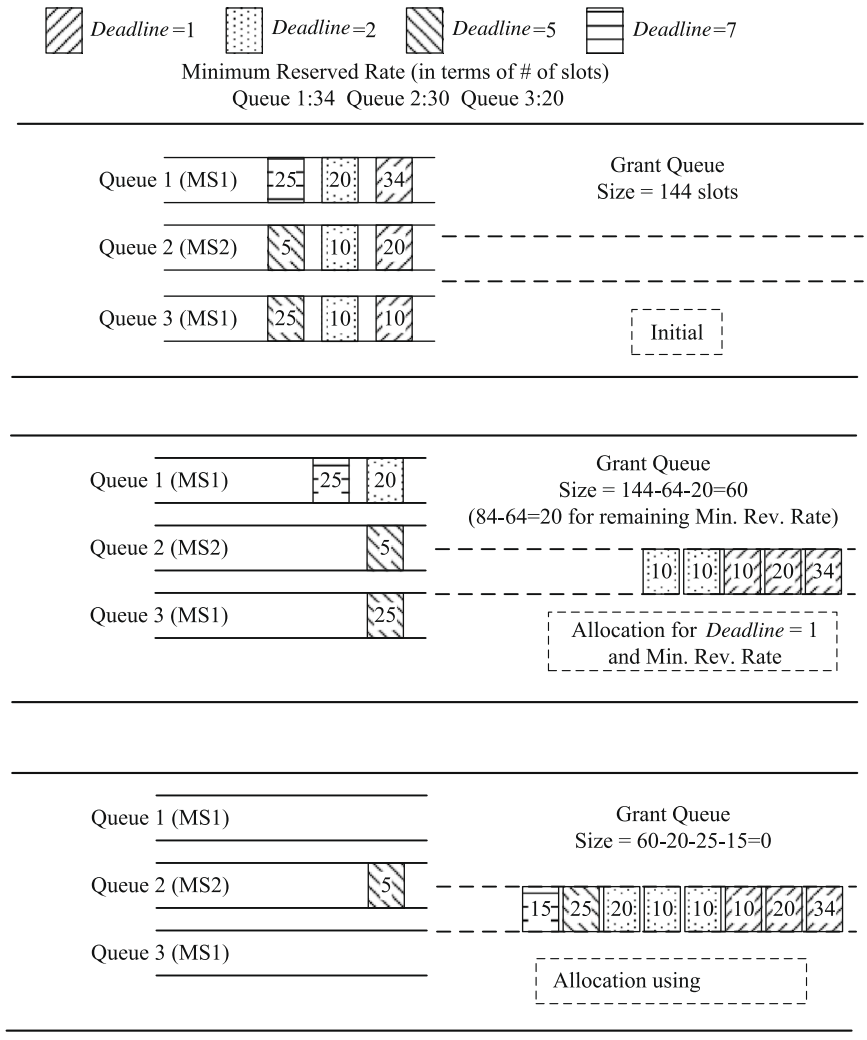

Fig. 4. Example of the urgency-based allocation in UL.

In the second phase, the HUF allocates slots to DL and UL requests, respectively. Take the UL as an example, while $\frac{36}{3} \times 12 = 144$ slots have been allocated in the first phase, in the second phase the bandwidth is reserved for requests whose deadline equals to one and also for the minimum reserved rate of the queues. This is accomplished by $144 - (34 + 20 + 10) - (0 + 10 + 10) = 60$ slots, since the HUF only needs to allocate additional $34 - 34 = 0$ slot for Queue 1, $30 - 20 = 10$ slots for Queue 2, and $20 - 10 = 0$ slot for Queue 3. Therefore, 60 slots are left to be allocated to queues according to their *average-U-factor*s, in which the queue of the largest *average-U-factor* is served first. As shown in Fig. 4 in which priority of each queue is configured to 0, the *average-U-factor* of all queues are calculated as $(\frac{20\times(0+1)}{2} + \frac{25\times(0+1)}{7})/2 \cong 6.79$, $\frac{5\times(0+1)}{5} = 1$, and $\frac{25\times(0+1)}{5} = 5$ for Queue 1, 2, and 3, respectively. Thus, the HUF selects the head-of-line request in Queue 1 to serve first, and recalculates the *average-U-factor* of Queue 1 as $\frac{25\times(0+1)}{7} \cong 3.57$. Similar procedures are executed until the sub-frame is fulfilled. Finally the HUF calculates the total number of slots each queue has just gained which are $34 + 20 + 15 = 69$, $20 + 10 = 30$, and $10 + 10 + 25 = 45$ for Queue 1, 2, and 3, respectively, and then grant them to every MS, namely $69 + 45 = 114$ for MS1, and 30 for MS2. Finally, the deadline of all requests is decreased by one before entering the next frame.

4. Evaluation results

The HUF algorithm is evaluated using the OPNET simulator with the WiMAX module developed by the INTEL Corp. The evaluation scenarios cover the MCS awareness, latency-aware dynamic adjustment, latency guarantee, and fairness in service classes. Each scenario considers a set of algorithms supporting certain functionality. For example, the DFPQ and MLWDF are involved when discussing the latency guarantee, while the DFPQ, TPP and UPS are considered for the fairness. Furthermore, only the rtPS and BE are involved in the evaluation because the UGS as well as ertPS is granted with fixed bandwidth, and the nrtPS differs from the BE merely in the priority.

4.1. Simulation environment

The simulation topology is depicted in the Fig. 5. A number of MSs and a BS are connected via a gateway to a video conference endpoint and an FTP server.

The video conference application used in the simulation has variable packet size and is constrained by the latency requirement.

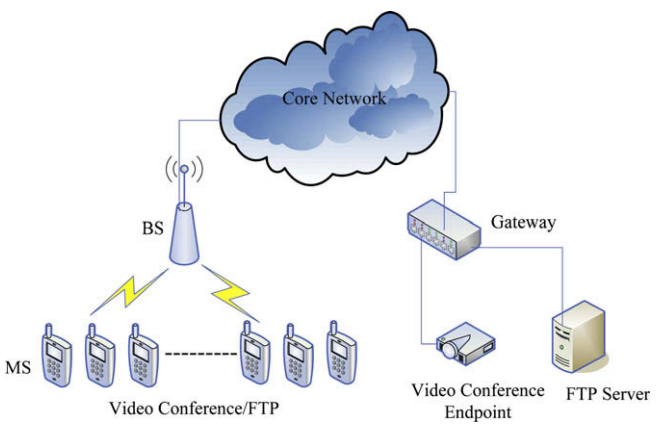

Fig. 5. Simulation topology.

Table 4
(a) System profile and (b) application parameters in the simulation.

System parameter		DL	UL
(a)			
System bandwidth		1.5 MHz	
FFT size		1024	
Frame duration		5 ms	
Useful symbol Time ($T_b = 1/f$)		60 μs	
Guard time ($T_g = T_b/8$)		7 μs	
OFDMA symbol duration ($T_s = T_b + T_g$)		67 μs	
Sub-channels		10	12
Number of slot per sub-channel		1	1
Number of symbols per slot		2	3
Application	**Parameter**		
(b)			
Video conference	Frame size: – Lognormal distribution – Average: 4.9 bytes – Standard deviation: 0.75 bytes [12] Frame inter-arrival time: $\frac{1}{30}$ s		
FTP	Requested file size: 200 Kbytes inter-request time: 30 s		

The WiMAX system profile [11] and application parameters are summarized in Table 4(a) and (b), respectively.

4.2. Modulation-aware allocation

Whenever the MCS is changed due to interferences, for consistent video conferencing quality the data rate of MSs is sustained by granting each of them adapted number of slots. Table 5 depicts the setup of the modulation awareness test of HUF, in which two MSs whose MCSs change along with time, are involved. From Fig. 6 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. Notably the system is not stable during the first 10 s because of performing Network Entry.

4.2.1. Latency-aware dynamic DL/UL adjustment

Dynamic DL/UL adjustment considering the latency requirement not only maximizes the link utilization but retains the quality of real-time applications. In this section, we evaluate the latency-aware adjustment supported by the HUF and compare its performance with the static approach and TPP. Six MSs are dedicated to downloading files using FTP with BE while an increasing number of MSs performing video conferencing with rtPS are adopted to enlarge the link load. Profiles of the applications are configured according to Table 4. Throughput and violation rate are investigated and shown in Fig. 7. Violation rate, defined as the ratio of the number of packets whose delay exceeds the maximum latency requirement to the number of all packets, is used to judge whether the adjustment is latency-aware.

As depicted in Fig. 7(a), the throughput of dynamic adjustment, whether using TPP or HUF, is about 7% higher than the static adjustment when overloaded with 41 MSs. This is due to the fact that the former dynamically exploits the bandwidth according to the DL and UL traffic loads, while the latter tends to abuse link resources because of not concerning the actual requirement. Fig. 7(b) shows that the degraded throughput of static adjustment contributes to the increased violation rate. Although the TPP has similar throughput to HUF, its violation rate is considerably higher than that of HUF, whose rate is close to zero. This is because the TPP decides the DL/UL allocation simply by considering their loads, while the HUF further reserves bandwidth for requests that must be served in the current frame.

Table 5
Setup of the modulation awareness test. The MCS changes along with time.

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

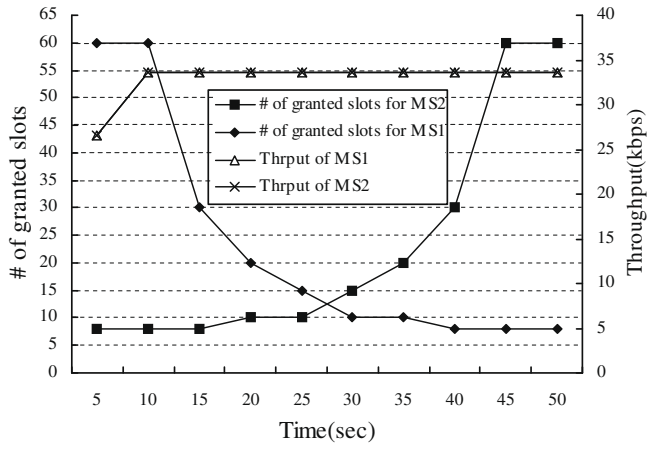

Fig. 6. Modulation-aware allocation: throughput is sustained as the MCS changes.

4.3. Latency guarantee with different requirements

We compare the latency-awareness of the proposed algorithm with the MLWDF, which is throughput-optimal and considers the waiting time of head-of-line packet to meet the latency quarantee, and with the DFPQ which uses EDF [8] for rtPS to satisfy the requirement. The evaluation scenario involves two flows of the video conference application referencing to Table 4(b). Among the QoS parameters of the two flows presented in Table 6, only the maximum latency is configured differently to 50 and 150 ms, respectively. The load of the link is increased by simultaneously increasing the input of the flows.

While throughput and average latency are the general criteria to evaluate the performance of a bandwidth algorithm, the violation rate is taken into account to estimate the satisfaction with different latency requirements, as discussed in Fig. 8 involving three algorithms. The criteria of the evaluation are throughput, average latency of packets and violation rate. The throughput and average

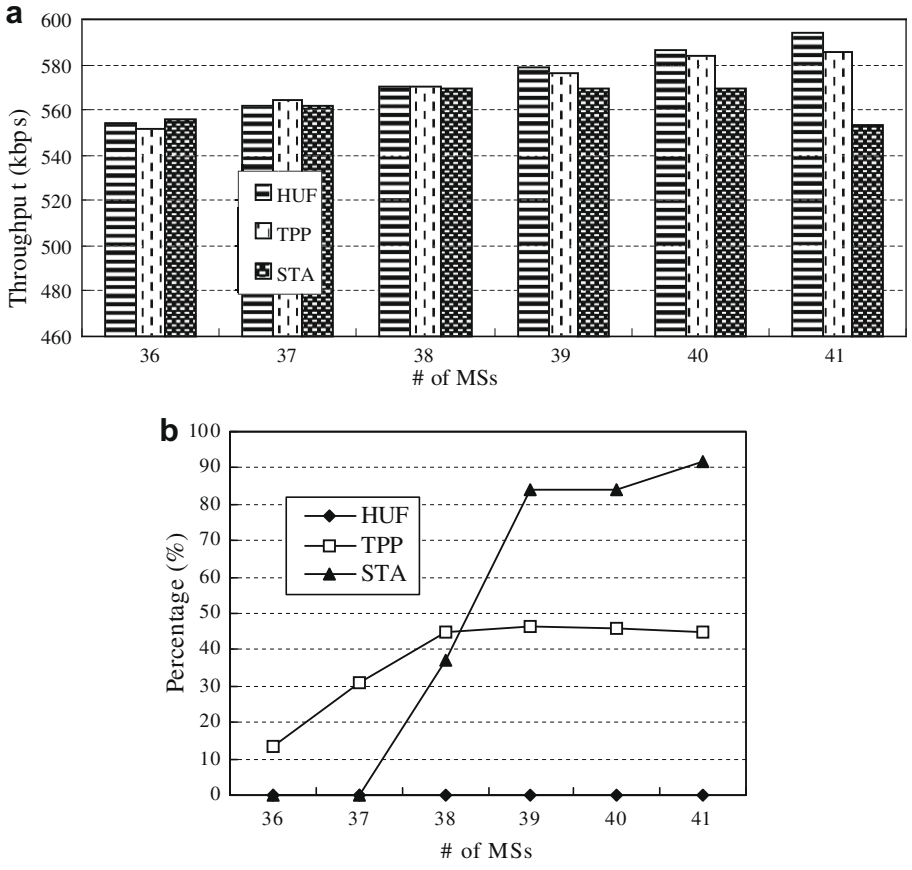

Fig. 7. (a) Throughput and (b) violation rate of three different algorithms after DL/UL adjustment.

Table 6
QoS parameters of two traffic flows.

QoS parameter	Flow #1	Flow #2
Service class	rtPS	rtPS
Minimum reserved rate (bps)	2400	2400
Maximum sustain rate (bps)	1,000,000	1,000,000
Maximum latency (ms)	50	150
Polling time (ms; specific to the rtPS class)	20	20

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. Fig. 8 discusses the throughput as well as the latency of three algorithms. From Fig. 8(a) we can observe that generally the throughput increases as more MSs participate in. However, performance of the DFPQ starts to degrade when the number of MSs reaches 32. This is because the EDF, which is an optimal scheduling algorithm in resource sufficient environment, malfunctions when overloaded [13]. The corresponding average latency in Fig. 8(b) is thus found to exceed 1000 ms suddenly from 32 MSs in DFPQ. The throughput is similar between the MLWDF and HUF, though the average latency differs noticeably when number of MSs approaches 30 since the MLWDF only considers the waiting time of the head-of-line packet, resulting in high average latency when heavily loaded. The HUF achieves high throughput while retaining low average latency.

Fig. 8(c) further examines the violation of the three algorithms in latency. Even when the number of MSs comes to 34, the HUF has no violation of the maximum latency being 50 and 150 ms. Nevertheless, the violation rate of MLWDF grows drastically when 28

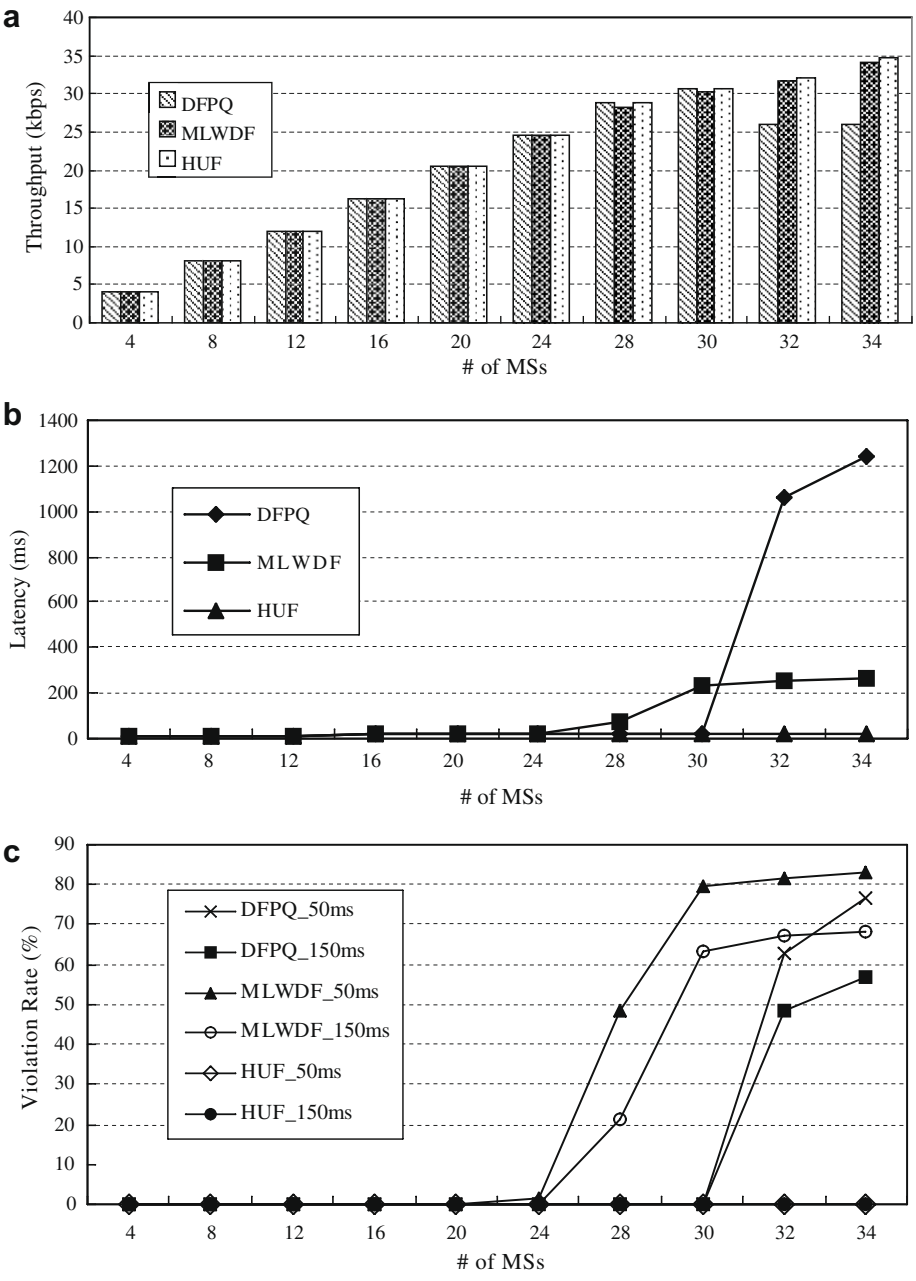

Fig. 8. (a) Throughput, (b) average latency and (c) violation rate of three different algorithms.

MSs are involved and is close to 70% and 80%, respectively, when maximum latency is configured to 150 and 50 ms and 34 MSs are present. This indicates that considering the head-of-line packet's waiting time may not be sufficient to guarantee the latency requirement. The DFPQ has a violation rate of 58% for 50 ms and 78% for 150 ms for 34 MSs resulted from the degraded throughput.

4.4. Fairness

A bandwidth allocation algorithm is said to be fair if the difference in normalized services received by different flows in the scheduler is bounded [8]. In the evaluation which compares the fairness of the HUF with DFPQ, TPP and UPS, two sets of MSs are involved with one performing rtPS-based video conferencing and the other uploading files via BE-based FTP. The application profiles are shown in Table 4(b) while the parameters of service classes are presented in Table 7.

The fairness between rtPS and BE can be formulated as

$$Fairness_{r\_b} = \left| \frac{Th_{rtPS}}{S_{rtPS}} - \frac{Th_{BE}}{S_{BE}} \right| \text{ [8]}, \quad (7)$$

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. The results are depicted in Fig. 9, in which small values suggest fair allocation.

Fig. 9(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 determining an appropriate maximum sustained rate for all service classes is not trivial. Thus, if the maximum sustained rate is not configured properly, the fairness suffers. Fig. 9(b) further explains the results. As shown in the figure, all approaches allocate fairly, namely 17% for rtPS and 83% for BE, when four MSs are employed. However, UPS and DFPQ start to distribute excessive number of slots to rtPS for eight MSs owing to its high priority, resulting in the starvation of BE. Contrastively, the HUF is quite fair even when 16 MSs are involved. TPP behaves similarly to the HUF, but becomes much unfair when heavily loaded because it tends to grant excessive slots to service classes.

Table 7
Parameters of the rtPS and BE.

QoS parameter	Type I	Type II
Service class	rtPS	BE
Minimum reserved rate (bps)	2400	2400
Maximum sustain rate (bps)	1,000,000	1,000,000
Maximum latency (ms)	50	N/A
Polling time (ms)	20	N/A

5. Conclusions and future work

This work aims at designing an integrated bandwidth allocation algorithm for a WiMAX BS in order to provide (1) dynamic down-

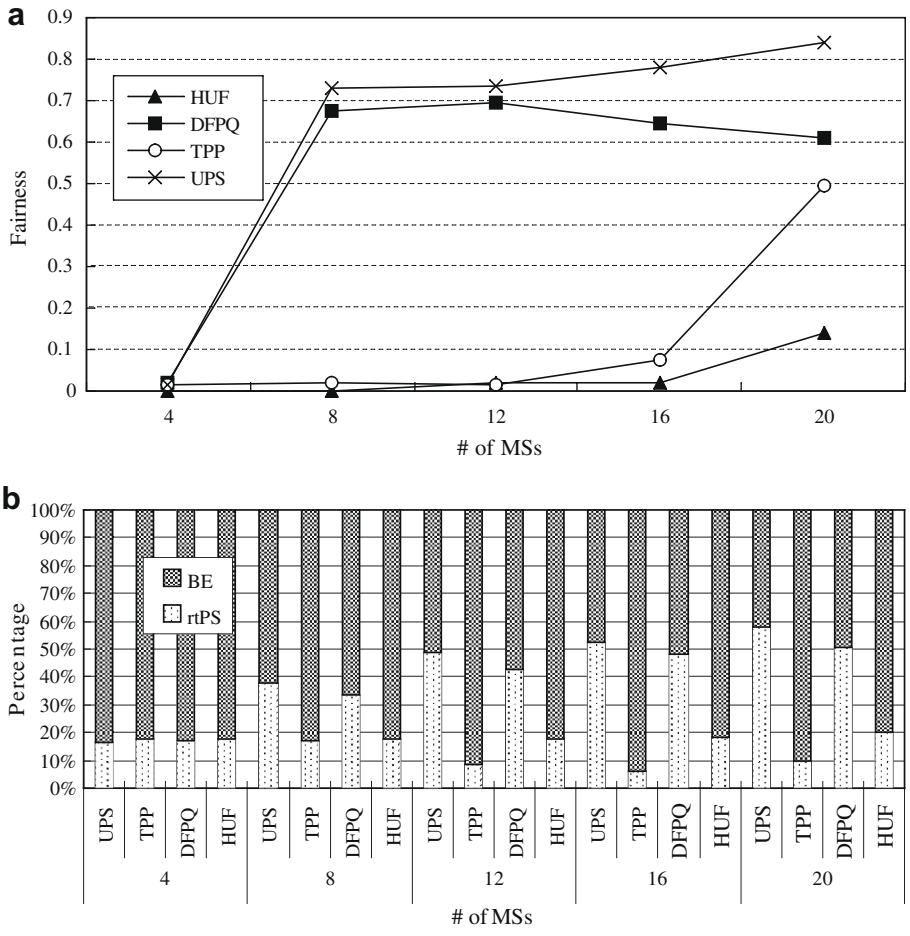

Fig. 9. (a) Fairness and (b) percentage of granted slots for rtPS and BE of four algorithms.

link/uplink adjustment, (2) latency guarantee for real-time applications, and (3) service differentiation and fairness among all service classes. Moreover, the MCS is incorporated to be adaptive to the link status between MSs and BS to lessen the impact from long distance and interference. The GPSS is preferred not only to comply with the standard but also to enable MSs to flexibly utilize the bandwidth.

The HUF is proposed to achieve the goals. 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 which considers latency, priority and fairness. A data/request with a deadline equaling to one needs to be served immediately, while others' urgency is calculated and indicated as the *U-factor*. In the dynamic DL/UL sub-frame determination, the HUF firstly reserves bandwidth for (1) data/requests whose deadline equals to one and (2) the minimum reserved rate of each service flow, and then proportionates the remaining bandwidth for DL/UL according to the remnant non-urgent data/requests. After that the head-of-line data/request of a queue with the largest *average-U-factor*is allocated, repeatedly, until the sub-frame is fulfilled. Finally, each MSs obtains grant from its own service queues.

Simulation result indicates that the quality is retained as the MCS adapts owing to the link quality. For dynamic adjustment, we show the throughput is as satisfactory as TPP and is 7% better than the static adjustment, and the violation rate is significantly alleviated by 42% and 80% compared to the TPP and static adjustment, respectively. The HUF outperforms the DFPQ by 25% in throughput when overloaded, and incurs no latency violation when the load is within system capacity. Finally, the HUF is observed to be fairer than the UPS, DFPQ and TPP and, unlike the TPP, it avoids inappropriate grant for rtPS.

Though HUF is relatively tolerant to overloaded situations, as a future direction we plan to develop admission control schemes to ease the degradation in throughput and fairness. Besides, while latency guarantee and fairness are concerned in BSs, a bandwidth allocation algorithm for MSs is also demanded to schedule appropriately the granted bandwidth.

References

[1] IEEE 802.16 Working Group, Air Interface for Fixed Broadband Wireless Access Systems, June 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, February 2006.
[4] IEEE 802.11 Working Group, Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications, September 1999.
[5] W.M. Yin, C.J. Wu, Y.D. Lin, Two-phase Minislot Scheduling Algorithm for HFC QoS Services Provisioning, GLOBECOM, November 2001.
[6] M. Andrews et al., Providing quality of services over a shared wireless link, IEEE Communication Magazine (February) (2001) 150–154.
[7] K. Wongthavarawat, A. Ganz, IEEE 802.16 Based Last Mile Broadband Wireless Military Networks with Quality of Service Support, MILCOM, October 2003.
[8] 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.
[9] Y.N. Lin, S.H. Chien, Y.D. Lin, Y.C. Lai, M. Liu, in: Maode Ma (Ed.), Dynamic bandwidth allocation for 802.16e-2005 MAC, Current Technology Developments of WiMax Systems, Springer, 2007.
[10] A. Sayenko, O. Alanen, J. Karhula, T. Hamalainen, Ensuring the QoS requirements in 802.16 scheduling, in: ACM MSWiM'06, October 2006.
[11] Mobile WiMAX Part I, A technical overview and performance evaluation, in: WiMAX Forum, April 2006.
[12] D.P. Heyman, A. Tabatabai, T.V. Lakshman, Statical analysis and simulation study of video teleconference traffic in ATM networks, IEEE Transactions on Circuits and Systems for Video Technology 2 (1) (1992) 49–59.
[13] C. Lu, J.A. Stankovic, G. Tao, S.H. Son, Design and evaluation of a feedback control EDF scheduling algorithm, in: Real-Time Systems Symposium, 1999.