

國立交通大學

網路工程研究所

碩士論文

延遲與調變感知的WiMAX

基地台動態頻寬分配演算法



Latency and Modulation Aware Dynamic

Bandwidth Allocation Algorithm for WiMAX Base Stations

研究生：吳哲文

指導教授：林盈達 教授

中華民國九十六年六月

延遲與調變感知的 WiMAX 基地台動態頻寬分配演算法

Latency and Modulation Aware

Dynamic Bandwidth Allocation Algorithm for WiMAX Base Stations

研究生：吳哲文

Student: Che-Wen Wu

指導教授：林盈達

Advisor: Dr. Ying-Dar Lin

國立交通大學

網路工程研究所



Submitted to Institutes of Network Engineering
College of Computer Science
National Chiao Tung University
in partial Fulfillment of the Requirements
for the Degree of
Master
in

Computer Science and Engineering

June 2007

HsinChu, Taiwan, Republic of China

中華民國九十六年六月

延遲與調變感知的 WiMAX 基地台動態頻寬分配演算法

學生：吳哲文

指導教授：林盈達

國立交通大學網路工程研究所

摘要

基於 802.16e-2005 標準的 WiMAX 行動系統宣稱在高速行動無線網路下可支援高速率且長距離的傳輸。然而連線品質受到長距離及空氣干擾而不穩定，使得及時應用程式面臨嚴峻的考驗。因此，一個適當的頻寬分配演算法同時滿足及時應用程式延遲要求且充分利用可用的頻寬來提供差異式服務和公平性是需要。在本論文中提出一個藉由考慮調適調變及編碼比率與需求急迫性的最高急迫優先演算法來克服上述的問題。在上下行頻寬的動態調節中，分別優先保留最急迫需求的頻寬，接著根據上下行個別不急迫的需求量比例地分配剩餘頻寬。在上下行獨自分配頻寬給行動裝置階段，則利用一新穎的係數來決定每個需求的急迫性，以最急迫優先的分配方式來反映出延遲保證與公平性。實驗模擬證實最高急迫優先演算法具備調變感知，同時保證即時串流的延遲而沒有超過其要求且維持跟其他演算法一樣的高流量表現。

關鍵字：頻寬分配，WiMAX，延遲，調變

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

Student: Che-Wen Wu

Advisor: Dr. Ying-Dar Lin

Institutes of Network Engineering

National Chiao Tung University

Abstract

The mobile WiMAX systems based on IEEE 802.16e-2005 provides high data rate for the mobile wireless network. However, the link quality is frequently unstable owing to the long-distance and air interference, leading to the crucible of real-time applications. Thus, a bandwidth allocation algorithm is required to satisfy (1) the latency requirement for real-time applications while supporting (2) service differentiation and (3) 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. This approach determines the downlink and uplink sub-frames by reserving the bandwidth for the most urgent requests and then proportionating the remaining bandwidth according to the non-urgent ones. Then, independently in the downlink and uplink, the HUF allocates bandwidth to every MS according to a pre-calculated U-factor which considers urgency, priority and fairness. The 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.

Keywords: bandwidth allocation, WiMAX, latency, modulation

Contents

| | |
|---|-----------|
| CHAPTER 1 INTRODUCTION..... | 1 |
| CHAPTER 2 BACKGROUND | 4 |
| 2.1 OVERVIEW OF THE WiMAX SYSTEM..... | 4 |
| 2.1.1 <i>PHY Layer Features.....</i> | <i>4</i> |
| 2.1.2 <i>MAC Layer with QoS.....</i> | <i>5</i> |
| 2.2 RELATED WORKS | 7 |
| 2.3 PROBLEM STATEMENT | 8 |
| CHAPTER 3 HIGHEST URGENCY FIRST..... | 9 |
| 3.1 OVERVIEW OF THE ALGORITHM..... | 9 |
| 3.2 DETAILED PROCEDURES OF THE ALGORITHM | 10 |
| 3.2.1 <i>Data/Request Translation and Deadline Determination.....</i> | <i>10</i> |
| 3.2.2 <i>First Phase: DL/UL Sub-frame Allocation</i> | <i>11</i> |
| 3.2.3 <i>Second Phase: Highest Urgency First Allocation.....</i> | <i>13</i> |
| 3.2.4 <i>Grant Bandwidth for Per MS.....</i> | <i>15</i> |
| 3.2.5 <i>Example</i> | <i>15</i> |
| CHAPTER 4 EVALUATION RESULTS..... | 19 |
| 4.1 SIMULATION ENVIRONMENT | 19 |
| 4.2 SIMULATION: EVALUATION AND RESULTS..... | 20 |
| 4.2.1 <i>Modulation-aware Allocation</i> | <i>20</i> |
| 4.2.2 <i>Latency-aware Dynamic DL/UL Adjustment</i> | <i>21</i> |
| 4.2.3 <i>Latency Guarantee with Different Requirements</i> | <i>23</i> |
| 4.2.4 <i>Fairness</i> | <i>26</i> |
| CHAPTER 5 CONCLUSIONS AND FUTURE WORKS | 29 |
| REFERENCE | 31 |

List of Figures

| | |
|--|----|
| FIG. 1. STRUCTURE OF A WiMAX OFDMA-TDD FRAME | 5 |
| FIG. 2. SCHEDULING FLOW AND QoS WITHIN BS AND MSS. | 7 |
| FIG. 3. PROCEDURE OF THE HIGHEST URGENCY FIRST (HUF). | 10 |
| FIG. 4. PSEUDOCODE OF THE FIRST PHASE OF HUF. | 13 |
| FIG. 5. PESUDOCODE OF THE SECOND PHASE OF HUF. | 15 |
| FIG. 6. THE EXAMPLE OF REQUESTED SLOTS IN ALL QUEUES IN UL..... | 18 |
| FIG. 7. SIMULATION TOPOLOGY. | 19 |
| FIG. 8. MODULATION-AWARE ALLOCATION: THE THROUGHPUT IS KEPT WHENEVER THE MCS IS CHANGED. | 21 |
| FIG. 9. A) THROUGHPUT AND B) VIOLATION RATE OF THREE DIFFERENT ALGORITHMS AFTER DL/UL ADJUSTMENT. | 23 |
| FIG. 10. A) THROUGHPUT, B) AVERAGE LATENCY AND C) VIOLATION RATE OF THREE DIFFERENT ALGORITHMS. | 26 |
| FIG. 11. A) FAIRNESS AND B) GRANTED SLOTS FOR RTPS AND BE OF FOUR ALGORITHMS..... | 28 |



List of Tables

| | |
|--|----|
| TABLE 1. SLOT SIZES OF DIFFERENT MCSs IN WiMAX..... | 5 |
| TABLE 2. SERVICE CLASSES AND THE CORRESPONDING QoS PARAMETERS..... | 6 |
| TABLE 3. A) DL/UL REQUESTED SLOTS AND B) SYSTEM PROFILE IN THE EXAMPLE..... | 17 |
| TABLE 4. A) SYSTEM PROFILE AND B) APPLICATION PARAMETERS IN THE SIMULATION. | 20 |
| TABLE 5. THE SCENARIO OF THE CHANGED MCS IN THE SIMULATION. | 21 |
| TABLE 6. THE QoS PARAMETERS OF THE TWO KINDS OF TRAFFIC FLOWS. | 24 |
| TABLE 7. THE PARAMETERS OF RTPS AND BE. | 26 |



Chapter 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, 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 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 [5]. In GPSS the BS grants requested bandwidth per MS rather than per connection so that the MS¹ 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

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.

interference. The MCS² 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, attempting to solve the above-mentioned considerations. The MLWDF [6] 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 [7] and DFPQ [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 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. 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. OFDMA-TDD, the most prevalent physical-layer technology for the WiMAX systems, has higher capacity of wideband owing to the technology of Orthogonal Frequency Division Multiple Access and more flexibility in the mobile environment than others. 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 indicates 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. Chapter 2 briefs the 802.16 PHY and MAC features and reviews related studies to justify our problems. Chapter 3 describes the detailed

² Below we use MCS to represent modulation types and channel coding scheme.

procedures of the proposed algorithm. Chapter 4 presents the simulation environments and evaluation results. Finally, Chapter 5 concludes this work with some future directions.

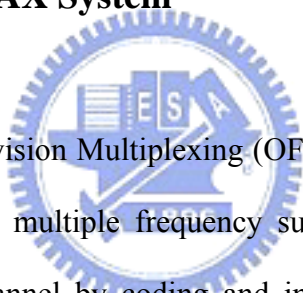


Chapter 2 Background

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 all decisions of transmissions from/to MSs, designating the importance of bandwidth allocation 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 subdivides 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. Orthogonal Frequency Division Multiple Access (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 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. 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 modulations and coding schemes (MCS) which decides the transmission data rate according to the link quality between the BS and MSs. Table 1 summarizes the bytes of a slot in all supported MCSs in WiMAX.

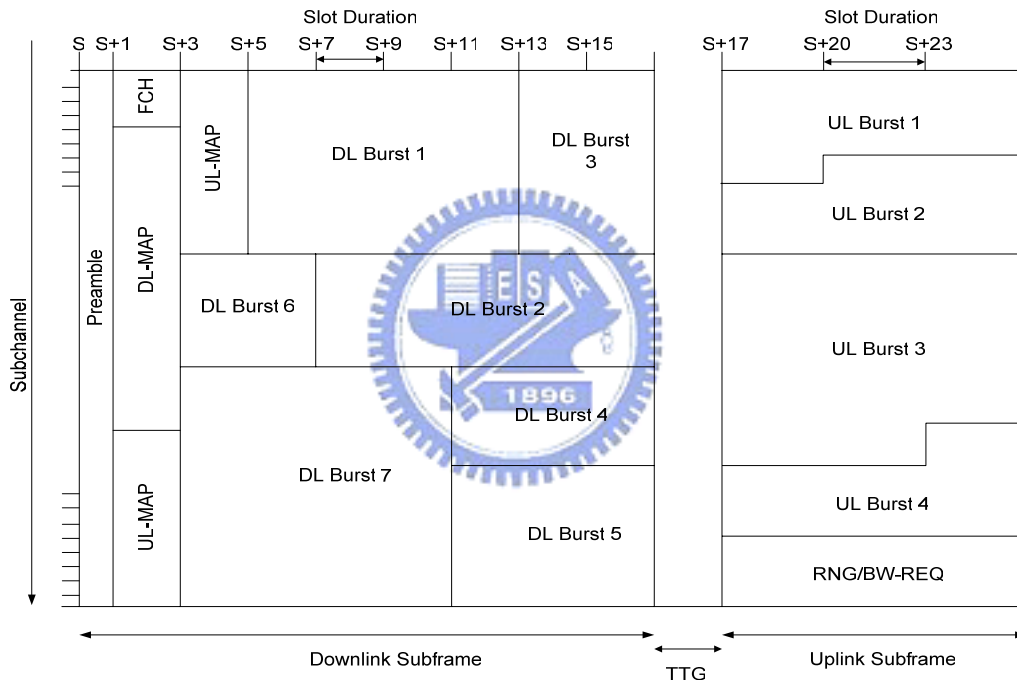


Fig. 1. Structure of a WiMAX OFDMA-TDD Frame.

Table 1. Slot sizes of different MCSs in WiMAX.

| Modulation | BSPQ | | | | QPSK | | | | 16QAM | | | | 64QAM | | | |
|------------|------|-----|-----|-----|------|-----|-----|-----|-------|-----|-----|-----|-------|-----|-----|-----|
| | 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 |

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 (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 enough bandwidth, the rtPS must be much concerned since it supports real-time applications having the maximum latency requirement and support variable packet sizes.

Table 2 summarizes the characteristics of those service classes.

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 service |

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

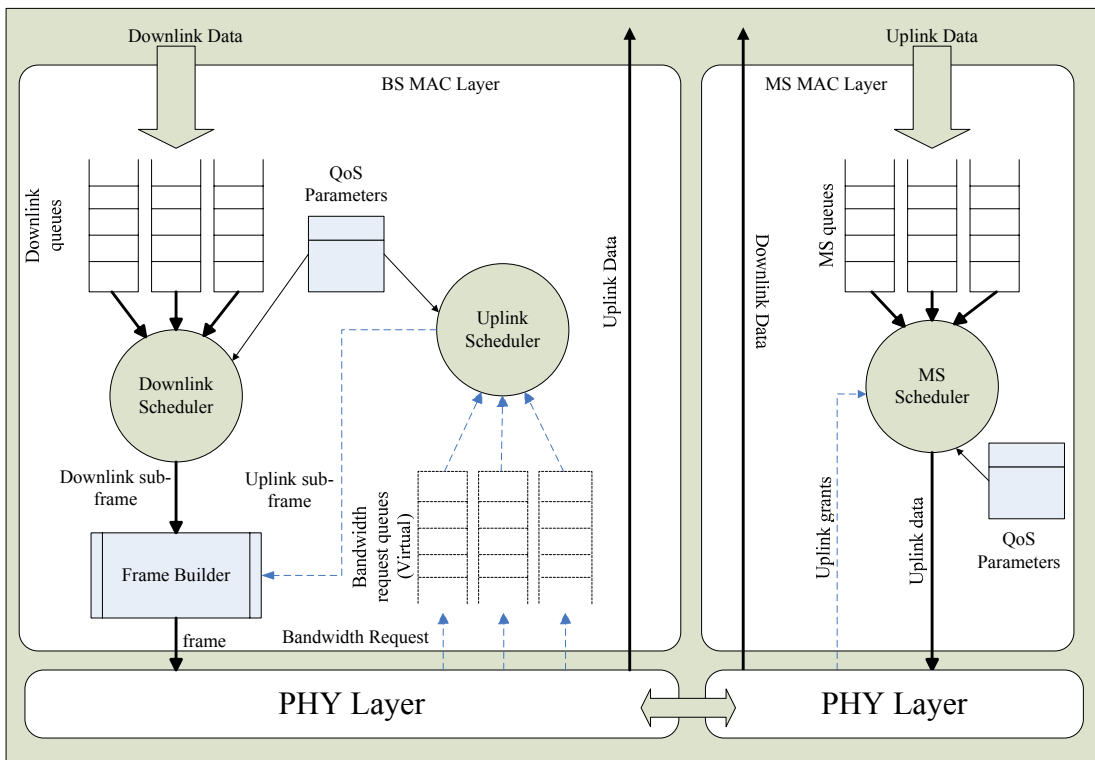


Fig. 2. Scheduling flow and QoS within BS and MSs.

2.2 Related Works

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. *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 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 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.

Chapter 3 Highest Urgency First

This chapter 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.

3.1 Overview of the Algorithm

The objective of the bandwidth allocation in WiMAX base stations is to fill up the dynamically adjusted DL/UL sub-frame in TDD mode in order to perform high throughput. Each sub-frame is further allocated to service queues of different QoS requirements such as latency guarantee, priority and fairness. Slots in the frame can carry different amount of data owing to the feature of adaptive modulation and coding schemes in PHY; the varying data rates may further affect how bandwidth allocation is performed. Based on the above characteristics, the *Highest Urgency First* (HUF) is proposed to well utilize the bandwidth. The *Urgency* parameter which considers three metrics, namely deadline, 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 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, HUF follows the GPSS by granting MSs the allocated bandwidth to each flow queue. The components and operations of the HUF algorithm are illustrated in Fig. 3 and explained in section 3.2.

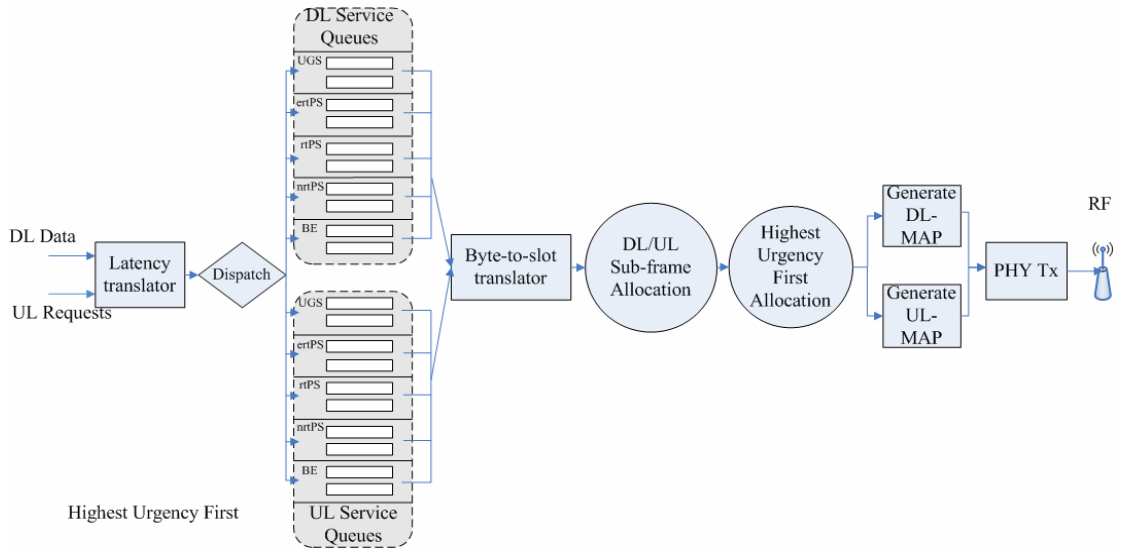


Fig. 3. Procedure of the Highest Urgency First (HUF).

3.2 Detailed Procedures of the 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 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 [11] 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.

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, 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:

- i. 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.
- 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} 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. Fig. 4 shows the pseudocode of the first phase.

```

whenever a frame start
//First Phase: decide the DL/UL sub-frame size
Update_MCS(); // update the modulation type and coding scheme
for Conn[CID].direct = UL {
  for every request in Queue[CID] {
    Request.slots = requesting_size/MCS_bytes; // translate the requesting size to slots in different MCS
    if Request.Deadline = 1 {
      UL_Rev_slots += Request.slots; // gather all reserved slots of Deadline = 1
      Request.D1[CID] += Request.slots; // all requested slots with Deadline = 1 of CID
    }
    else UL_Rem_slots += Request.slots; // gather all remaining requested slots
  }
  slots = Translate_Rmin_slots(Conn[CID].Rmin); // translate the min. reserved rate to slots
  slots -= Request.D1[CID]; // subtract the request with Deadline = 1 from min. rev. rate
  if slots > 0 {
    UL_Rev_slots += slots; // gather all reserved slots of min. reserved rate
    UL_Min_Rev_slots += slots; // the total remaining reserved slots for min. rev. rate
  }
}

for Conn[CID].direct = DL {
  for every Request in Queue[CID] {
    Request.slots = requesting_size/MCS_bytes; // translate the requesting size to slots in different MCS
    if Request.Deadline = 1 {
      DL_Rev_slots += Request.slots; // gather all reserved slots of Deadline = 1
      Request.D1[CID] += Request.slots; // all requested slots with Deadline = 1 of CID
    }
    else DL_Rem_slots += Request.slots; // gather all remaining requested slots
  }
  slots = Translate_Rmin_slots(Conn[CID].Rmin); // translate the min. reserved rate to slots
  slots -= Request.D1[CID]; // subtract the request with Deadline = 1 from min. rev. rate
  if slots > 0 {
    DL_Rev_slots += slots; // gather all reserved slots of min. reserved rate
    DL_Min_Rev_slots += slots; // the total remaining reserved slots for min. rev. rate
  }
}

// calculate the reserved symbols in DL/UL
UL_Rev_symbols = UL_slot_duration_symbols*ceil(UL_Rev_slots/#_UL_subchannel);
DL_Rev_symbols = DL_slot_duration_symbols*ceil(DL_Rev_slots/#_DL_subchannel);
// calculate the redundant reserved slots
UL_Red_slots = #_UL_subchannel - (UL_Rev_slots mod #_UL_subchannel);
DL_Red_slots = #_DL_subchannel - (DL_Rev_slots mod #_DL_subchannel);

// calculate the remaining symbols
Rem_symbols = Total_symbols - UL_Rev_symbols - DL_Rev_symbols;

// subtract the reserved slots of min. rev. rate from the remaining requested slots
DL_Rem_slots = (DL_Min_Rev_slots+DL_Red_slots) ;
UL_Rem_slots = (UL_Min_Rev_slots+UL_Red_slots) ;

// proportion the remaining symbols for DL/UL
x = (DL_Rem_slots*Rem_symbols) / (UL_Rem_slots*UL_slot_duration_symbols+DL_Rem_slots*DL_slot_duration_symbols);

DL_added_symbols = DL_slot_duration_symbols * ceil(x);
UL_added_symbols = Rem_symbols- DL_added_symbols;

// decide the DL/UL sub-frame size
UL_subframe_size = UL_Rev_symbols+UL_added_symbols;
DL_subframe_size = DL_Rev_symbols+DL_added_symbols;

```

Fig. 4. Pseudocode of the first phase of HUF.

3.2.3 Second Phase: 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:

- i. For each service flow, allocate bandwidth firstly to requests whose deadline equals to one and then to others until the minimum reserved rate is complemented.
- ii. 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

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

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

indicates the Urgency of the *i*th request in the flow and *n* represents number of requests. As shown in Eq. 6, the *U-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. Figure 5 presents the pseudocode of the above-mentioned procedure.

```

//Second Phase
Avail_slots_DL = (DL_subframe_size/DL_slot_duration_symbols)*#_DL_subchannel; // get the available slots in DL
Avail_slots_UL = (UL_subframe_size/UL_slot_duration_symbols)*#_UL_subchannel; // get the available slots in UL
if Avail_slots_UL > 0 {
    // allocate the bandwidth for that requests' Deadline is 1 first
    for Conn[CID].direct = UL {
        for every Request in Queue[CID] {
            if Request.Deadline = 1 {
                UL_Grant[CID] += Request.slots;
                Avail_slots_UL -= Request.slots;
                Remove(Request);
            }
        }
        if Avail_slots_UL <= 0 break;
    }
}
// allocate the bandwidth to SFs for satisfying the minimum reserved rate
for Conn[CID].direct = UL {
    slots = Translate_Rmin_slots(Conn[CID].Rmin) - UL_Grant[CID];
    if slots > 0 {
        UL_Grant[CID] += slots;
        if Avail_slots_UL >= slots Avail_slots_UL -= slots;
        else { slots = Avail_slots_UL; Avail_slots_UL = 0; }
        while slots > 0 {
            if HeadRequest(Queue[CID]).slots > slots { HeadRequest(Queue[CID]).slots -= slots; break; }
            else { slots -= HeadRequest(Queue[CID]).slots; Remove(HeadRequest); }
        }
    }
    if Avail_slots_UL <= 0 break;
}

// calculate the U-factor for requests and chose the maximum one to serve first
while Avail_slots_UL > 0 {
    for all Conn[CID].direct = UL {
        for every Request in Queue[CID] {
            Total_U_factor += ((Request.slots)*(Conn[CID].priority+1)/(Request.Deadline));
            Num_of_request++;
        }
        Queue[CID].avg_U_factor = Total_U_factor/Num_of_request;
    }

    for all Conn[CID].direct = UL {
        Max_CID = Max(Queue[CID].avg_U_factor);
    }

    UL_Grant[Max_CID] += HeadRequest(Queue[Max_CID]).slots;
    Remove(HeadRequest(Queue[Max_CID]));
    Avail_slots_UL -= HeadRequest(Queue[Max_CID]).slots;
}
}

if Avail_slots_DL > 0
do the same procedures as UL

```

Fig. 5. Pesudocode of the second phase of HUF.

3.2.4 Grant Bandwidth for Per 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 $Deadline = Deadline - 1$.

3.2.5 Example

This section elaborates an example of the HUF, in which the parameters and system profile are shown in Table 3. It is assumed that both DL and UL have four queues and the

minimum reserved rates are 24, 20, 20 and 20 slots for Queue 1, 2, 3 and 4, respectively. In the first phase, HUF decides the DL/UL sub-frame size. 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. Then, HUF proportionates the remaining symbols for the DL and UL by solving the equation (4) where DR is $50 + 50 + 20 - 20 = 100$ and UR is $40 + 30 + 25 - 20 = 75$. So, DL obtains additional $x = \frac{100 \times 29}{75 \times 3 + 100 \times 2} \cong 7$ slot durations equaling to $2 \times 7 = 14$ symbols and UL obtains additional $\frac{29 - 2 \times 7}{3} = 5$ 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.

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 $144 - (24 + 10 + 20 + 10) - (0 + 10 + 0 + 10) = 60$, since the HUF allocates $24 - 24 = 0$ for Queue 1, $20 - 10 = 10$ for Queue 2, $20 - 20 = 0$ for Queue 3, and $20 - 10 = 10$ for Queue 4. Therefore, 60 slots are left to be allocated to queues according to their *average-U-factor*, in which the queue of the largest *average-U-factor* is served first. As shown in Fig. 6 in which priority of each queue is configured to 0, the *average-U-factor* of all queues are calculated as $(\frac{15 \times (0+1)}{2} + \frac{25 \times (0+1)}{7})/2 \cong 5.54$, $\frac{5 \times (0+1)}{5} = 1$, $(\frac{5 \times (0+1)}{2} + \frac{10 \times (0+1)}{5})/2 = 2.125$, and $\frac{15 \times (0+1)}{5} = 3$ for Queue 1, 2, 3 and 4, 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

24 + 15 + 25 = 64, 10 + 10 = 20, 20 + 5 = 25, and 10 + 10 + 15 = 35 for Queue 1, 2, 3 and 4, respectively, and then grant them to every MS, namely 64 + 35 = 99 for MS1, 20 for MS2 and 25 for MS3. Finally, the deadline of all requests is decreased by one before entering the next frame.

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

(a)

| Direction | Type | Requested slots |
|------------------------------|---------------------|-----------------|
| UL (sum of all UL queues) | <i>deadline</i> = 1 | 64 |
| | <i>deadline</i> = 2 | 40 |
| | <i>deadline</i> = 5 | 30 |
| | <i>deadline</i> = 7 | 25 |
| | Min. Rev. Rate | 20 |
| DL (sum of all DL queues) | <i>deadline</i> = 1 | 90 |
| | <i>deadline</i> = 3 | 50 |
| | <i>deadline</i> = 4 | 50 |
| | <i>deadline</i> = 6 | 20 |
| | Min. Rev. Rate | 20 |

(b)

| PHY Parameter | Value |
|-----------------------------|-------|
| Symbols of a frame | 72 |
| Number of UL Sub-channels | 12 |
| Number of DL Sub-channels | 10 |
| UL Slots Duration (Symbols) | 3 |
| DL Slots Duration (Symbols) | 2 |

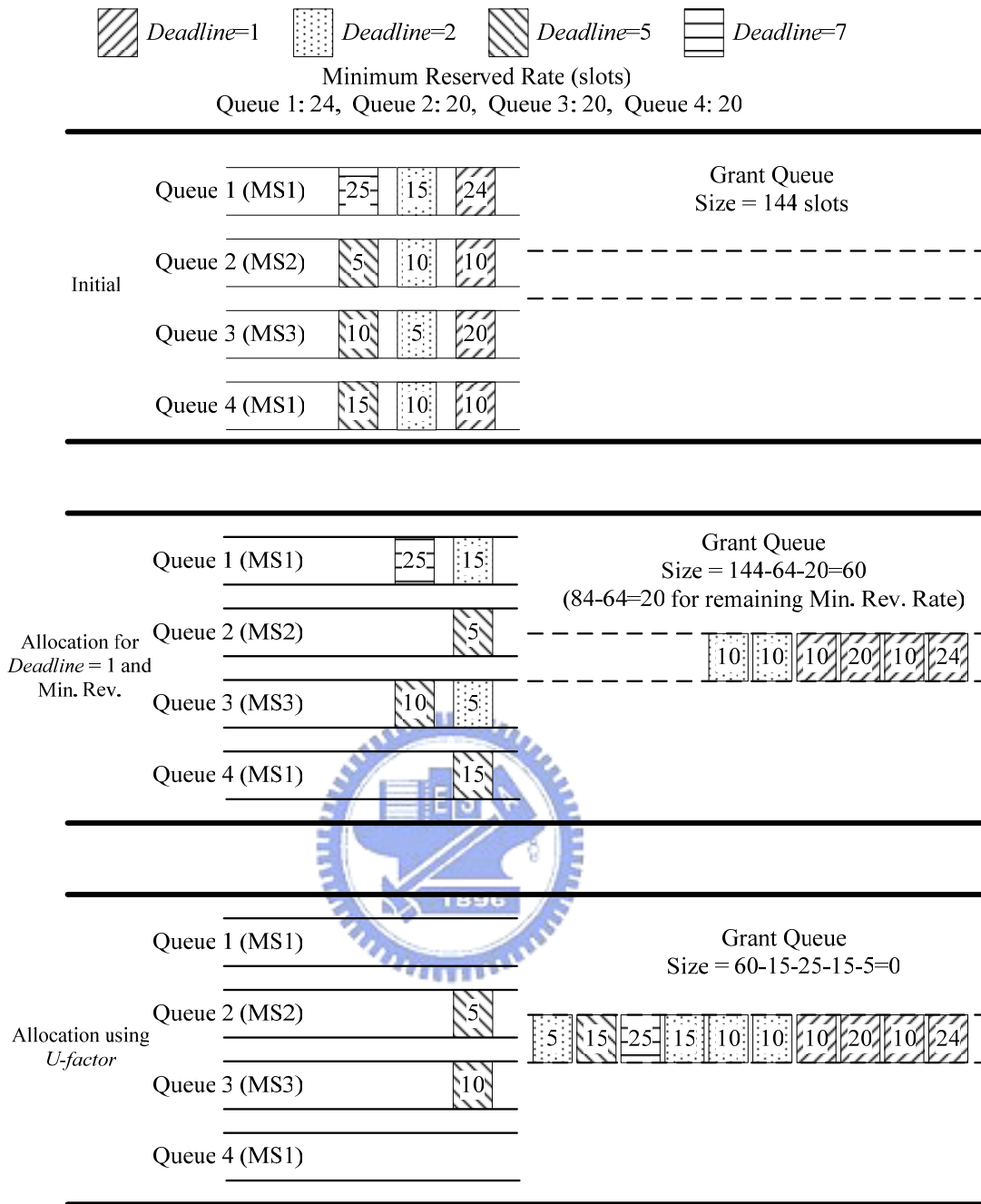


Fig. 6. The example of requested slots in all queues in UL.

Chapter 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. 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.

4.1 Simulation Environment

The simulation topology is depicted in the Fig. 7. A number of MSs and a BS are connected via a gateway to a video conference endpoint and an FTP server.

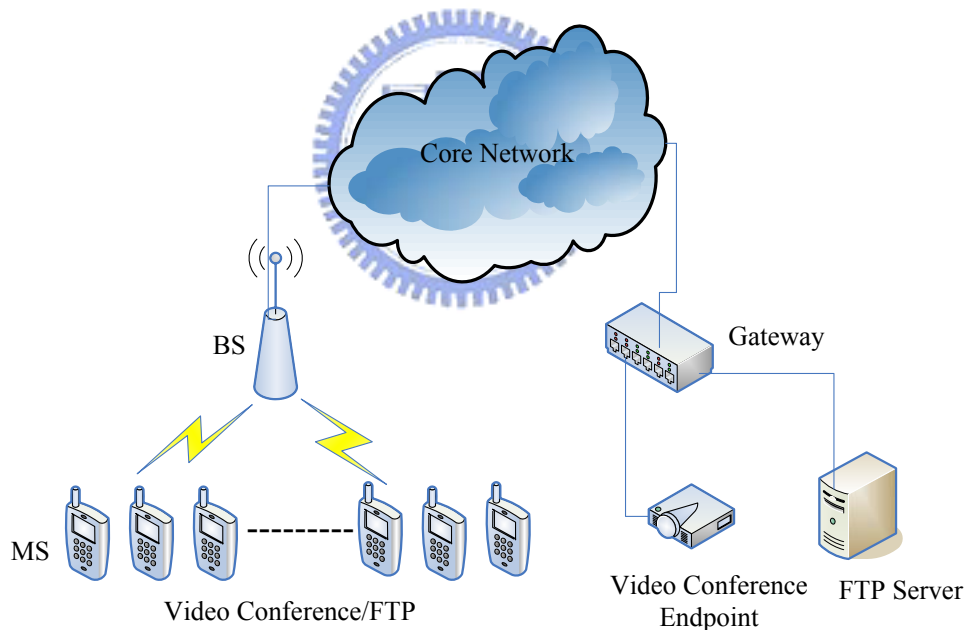


Fig. 7. Simulation Topology.

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

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

(a)

| 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 ($T_g = T_b/8$) | 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 |

4.2 Simulation: Evaluation and Results

This section itemizes the scenario and criteria of evaluation focusing on the modulation-aware allocation, latency-aware dynamic adjustment, latency guarantee with different requirements and fairness.

4.2.1 Modulation-aware Allocation

Whenever the MCS is changed due to various interferences, for consistent video conferencing quality the data rate of MSs must be sustained by granting each of them adapted number of slots. Table 5 depicts the modulation awareness of the HUF, in which two MSs whose MCS is changed along with time are involved. From Figure 8 we observe that though

modulation is changed constantly, the throughput is still kept at the same rate. The fewer bits a slot carries, the more slots are granted; similar behaviors occur otherwise

Table 5. The scenario of the changed MCS in the simulation.

| 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 |

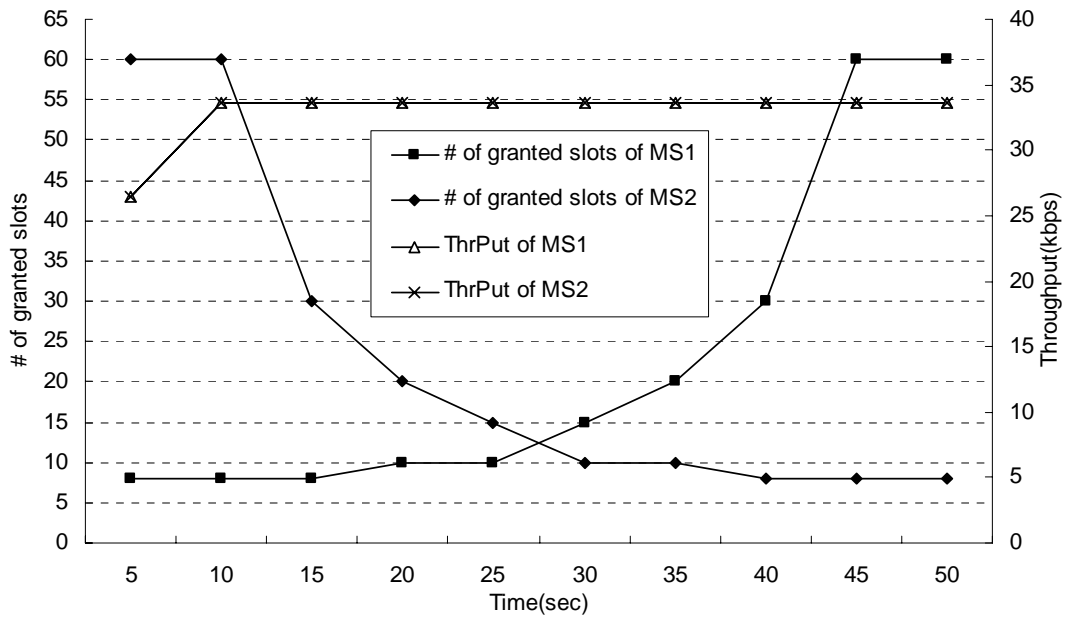


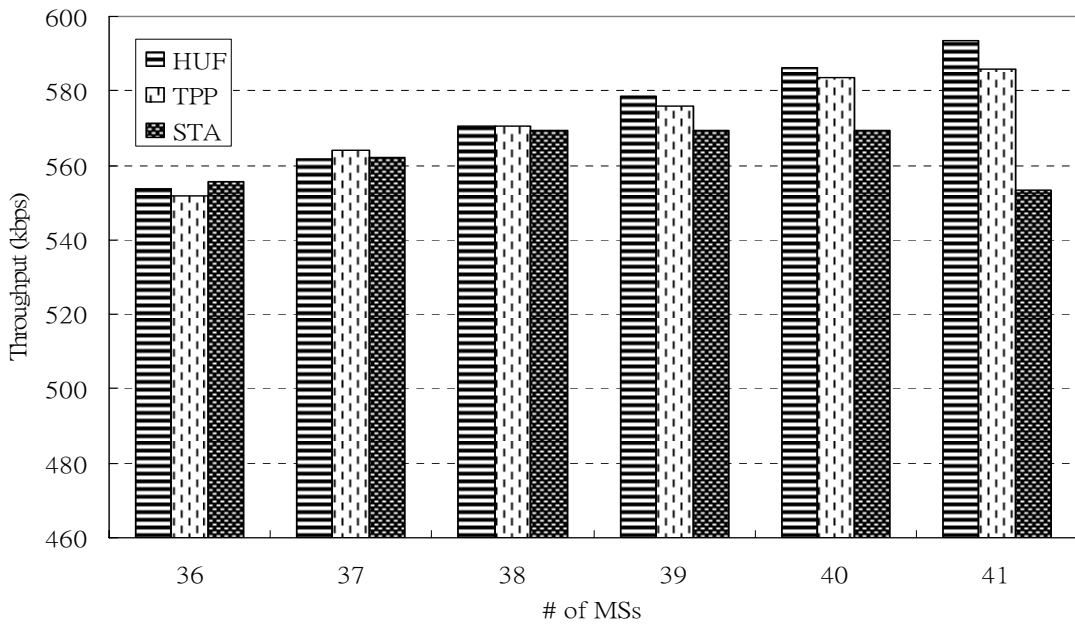
Fig. 8. Modulation-aware allocation: the throughput is kept whenever the MCS is changed.

4.2.2 Latency-aware Dynamic DL/UL Adjustment

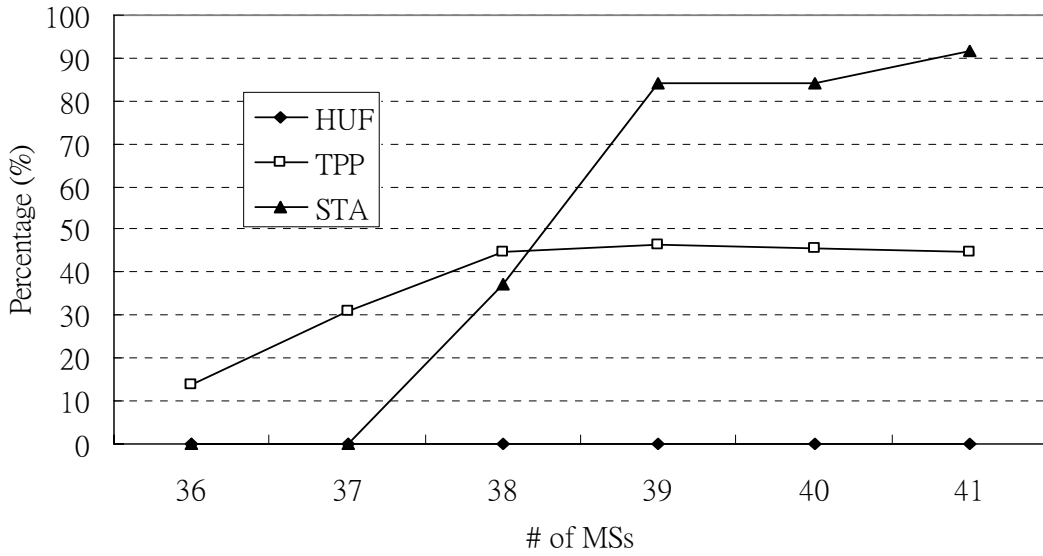
The dynamic adjustment of downlink and uplink sub-frame size is used to maximize the throughput of the link. Besides, the adjustment must take the requests with latency requirement into consideration for keeping the quality. In this section, we introduce the scenario of the evaluation for the latency-aware adjustment, and compare the performances by using static adjustment and dynamic adjustment based on TPP or the proposed algorithm HUF.

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 HUF and compare the performance with the static approach and the TPP. Six MSs are dedicated for 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. 9. Violation rate which means the ratio of the number of packets whose delay exceeds its maximum latency requirement to all number of packets is used to judge whether the adjustment is latency aware. An adjustment is said to be latency-aware if it considers the latency requirements to bring about low violation rate.



(a)



(b)

Fig. 9. a) Throughput and b) violation rate of three different algorithms after DL/UL adjustment.

As depicted in Fig. 9(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 waste link resources when the traffic loads contrast much to the static allocation. Figure 9(b) shows that the degraded throughput of static adjustment contributes to the increasing 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.

4.2.3 Latency Guarantee with Different Requirements

Latency guarantee in rtPS is critical for proper QoS. Though the requirement is different, the bandwidth allocation algorithm must guarantee and satisfy for them. In this section, 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 [8] for rtPS to satisfy the requirement. The evaluation scenario uses the video conference application referencing Table 4(b) based on two sets of QoS parameters used in rtPS presented in Table 6. 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.

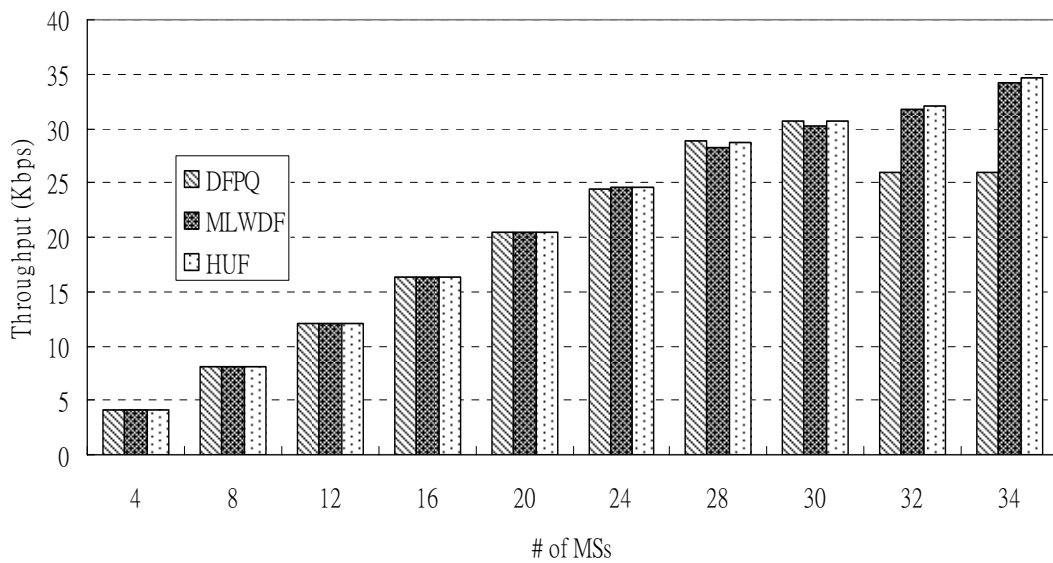
Table 6. The QoS parameters of the two kinds of traffic flows.

| 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 10 discusses the throughput as well as the latency of three algorithms. From Fig. 10(a) we can observe that generally the throughput increases as more MSs participate in. However, 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, degrades rapidly when overloaded [13]. The corresponding average latency in Fig. 10(b) is thus found to exceed 1000ms suddenly from 32 MSs in DFPQ. The throughput is similar between the MLWDF and HUF, though the average latency differs by 247ms, since the MLWDF only considers the head-of-line waiting time which results in high average latency when heavily loaded. The HUF achieves high throughput while retaining low average latency.

Figure 10(c) further examines the violation of the three algorithms in latency. Even when

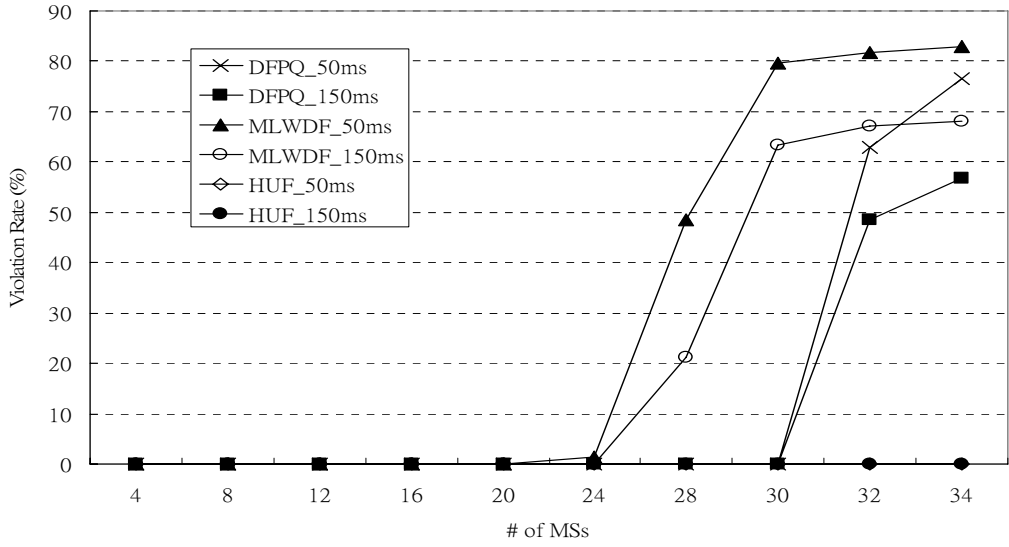
the number of MSs is up to 34, the HUF has no violation for the maximum latency being 50ms and 150ms. Nevertheless, the violation rate of MLWDF grows drastically when 28 MSs are involved and is close to 70% and 80%, respectively for maximum latency requirement being 150ms and 50ms when 34 MSs are present. This indicates that considering the head-of-line packet's waiting time may not be sufficient to guaranteeing the latency requirement. The DFPQ has a violation rate of 58% for 50ms and 78% for 150ms for 34 MSs resulted from the degraded throughput.



(a)



(b)



(c)

Fig. 10. a) Throughput, b) average latency and c) violation rate of three different algorithms.

4.2.4 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 point of the above description, 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 4(b) while the parameters of service classes are presented in Table 7.

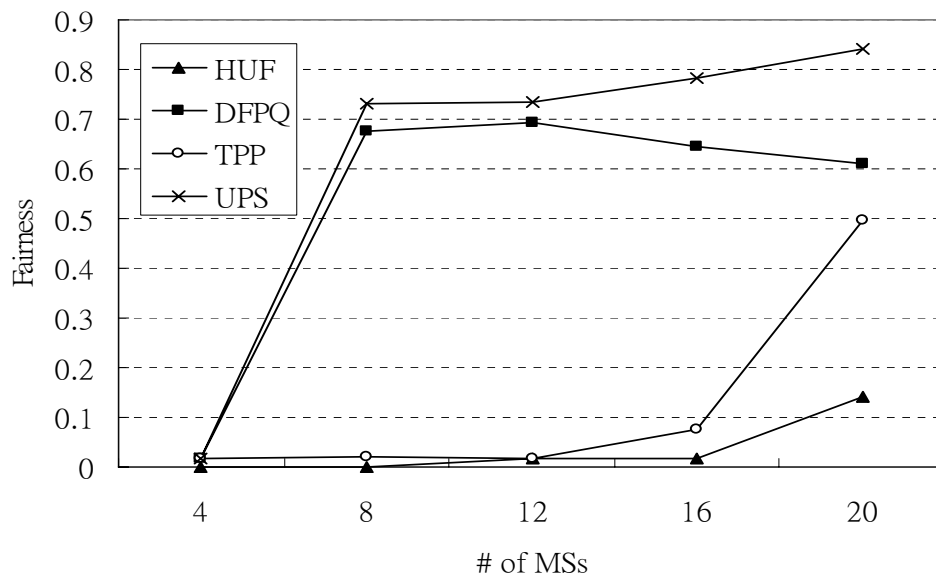
Table 7. The parameters of rtPS and BE.

| 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 |

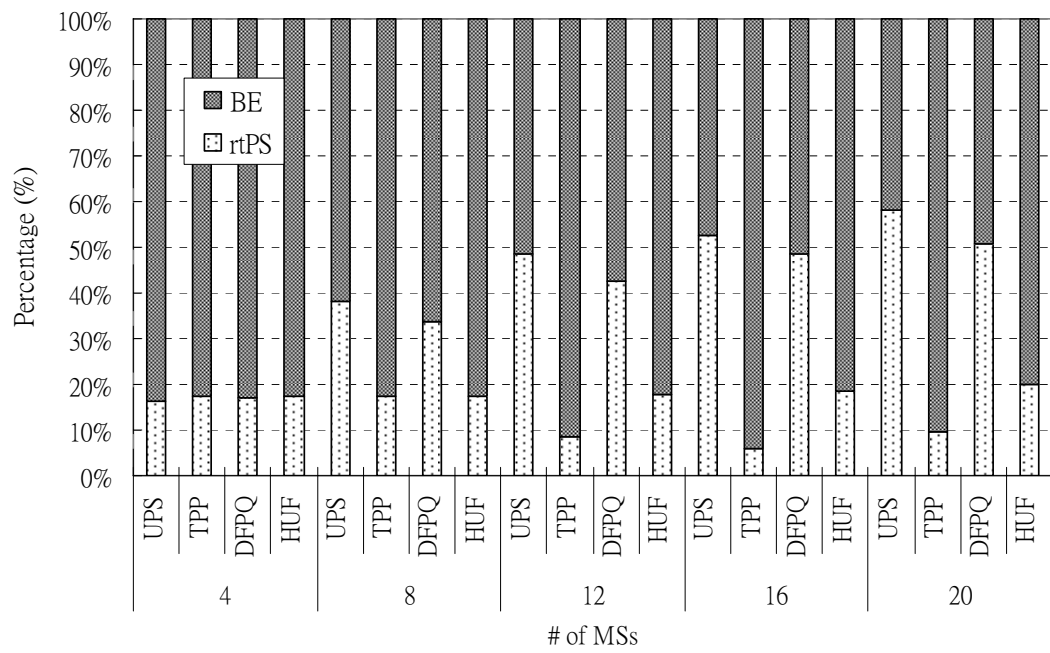
The fairness between rtPS and BE can be formulated as

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

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. 11, in which small values suggest fair allocation. Figure 11(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. Figure 11(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 8MSs because of higher priority, resulting in the starvation of BE. Contrastively, HUF is quite fair even when 16MSs are involved. TPP behaves similarly to the HUF, but becomes much unfair when heavily loaded because it tends the proportion leads to grant more slots to one of service classes which exceed the need.



(a)



(b)

Fig. 11. a) Fairness and b) granted slots for rtPS and BE of four algorithms.



Chapter 5 Conclusions and Future Works

This work aims at designing an integrated bandwidth allocation algorithm for a WiMAX BS in order to guarantee the latency requirement of real-time applications as well as service differentiation and fairness among all service classes. Dynamic downlink/uplink adjustment is also employed to well utilize the scarce wireless link. Since the mobility is supported in the WiMAX system in which link quality frequently changes due to long distance and interference, the modulation and coding schemes need to be adaptive to the link status between MSs and BS. Moreover, the Grant-Per-SS (GPSS) is preferred not only to comply with the standard but also to provide MSs the flexibility of domination.

The *Highest Urgency First* (HUF) is proposed to achieve the above goals. HUF translates the requested size to number of slots according to current MCS when every frame starts and uses the *Urgency* of data/request as criterion of allocation. A data/request with a deadline equaling to one is the most urgent and needs to be allocated immediately, while others' urgency is decided by 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 satisfying (1) and (2) for DL/UL allocation to queues, the head-of-line data/request of a queue with the largest *average-U-factor* is granted one by one until the sub-frame is fulfilled. Finally, each MSs obtains its 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 good as TPP and increases 7% compared with static adjustment, and the violation rate is better about 42% and 80% than TPP and static adjustment respectively. HUF outperforms the DFPQ by 25% in throughput when

overloaded, and incurs no latency violation within system capacity. Finally, we compare the fairness of UPS, DFPQ, TPP and HUF and observe fairness between rtPS and BE in HUF which 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 of the throughput and fairness. Besides, while latency guarantee and fairness are now concerned in BSs, bandwidth allocation algorithms for MSs are also demanded to schedule appropriately the granted bandwidth.



Reference

- [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] W. M. Yin, C. J. Wu, Y. D. Lin, "Two-phase Minislot Scheduling Algorithm for HFC QoS Services Provisioning," GLOBECOM, Nov. 2001.
- [6] M. Andrews et al., "Providing Quality of Services over a Shared Wireless link," IEEE Communication Magazine, pp. 150-154, Feb. 2001.
- [7] K. Wongthavarawat, A. Ganz, "IEEE 802.16 Based Last Mile Broadband Wireless Military Networks with Quality of Service Support," MILCOM, Oct. 2003.
- [8] J. Chen, W. Jiao, H. Wang, "A Service Flow Management Strategy for IEEE802.16 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, "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.
- [10] A. Sayenko, O. Alanen, J. Karhula, T. Hamalainen, "Ensuring the QoS Requirements in 802.16 scheduling," ACM MSWiM '06, Oct. 2006.
- [11] Mobile WiMAX Part I, "A Technical Overview and Performance Evaluation," 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 Trans. Circuits Sys. Video Tech., vol. 2, no. 1, Mar. 1992, pp.49-59.
- [13] C. Lu, J. A. Stankovic, G. Tao, S. H. Son, "Design and Evaluation of a Feedback Control EDF Scheduling Algorithm," Real-Time Systems Symposium, 1999.