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.



## 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

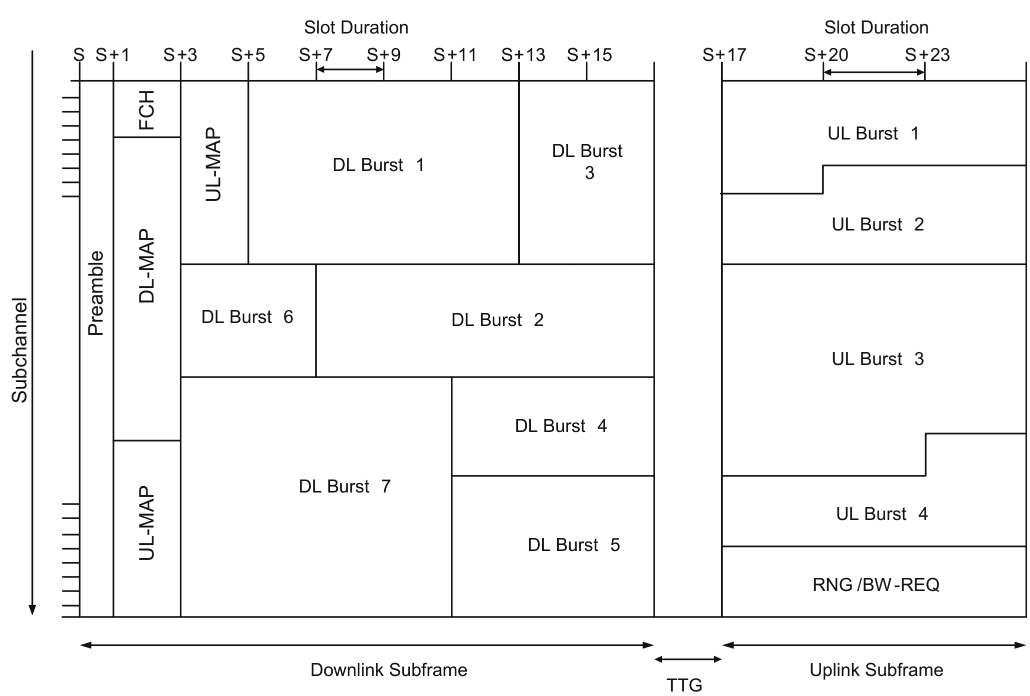


**Fig. 1.** Structure of a WiMAX OFDMA-TDD frame.