Adaptive Time-utility Function Scheme for Downlink Packet Scheduling in IEEE 802.16e/WiMAX Networks

JUI-CHI CHEN*
Department of Computer Science and Information Engineering, Asia University, Taiwan

ABSTRACT
WiMAX is a well-known broadband wireless access technology that supports multimedia transmission users. IEEE 802.16e has evolved into five quality of service (QoS) scheduling types for WiMAX uplink transmissions, but downlink transmission types have not yet been defined. To assist in this task, and to support efficient base station packet scheduling that takes QoS into account, I propose an adaptive time-utility function (A-TUF) packet scheduling scheme to improve downlink real-time (RT) traffic QoS and whole system throughput. Simulation results indicate that the proposed scheme is capable of providing lower RT packet delay and loss rate compared to conventional TUF-based algorithms.

Key words: IEEE 802.16e, WiMAX, packet scheduling, adaptive time-utility function.

1. INTRODUCTION

Worldwide Interoperability for Microwave Access (WiMAX) is a well-known broadband wireless access technology that features wide coverage, high bandwidth and long-distance transmission capabilities (Li, Qin, Low & Gwee, 2007). An IEEE 802.16 sub-committee has issued a series of WiMAX standards, including IEEE 802.16e (Mobile WiMAX) (IEEE 802.16 Working Group [IEEE 802.16WG], 2005). Regarding the WiMAX physical layer, orthogonal frequency-division multiple access (OFDMA) is now viewed as a potential key technology (Ali, Lee & Leung, 2007; Park, Cho & Bahk, 2008). OFDMA can adjust channel bandwidth and allocate corresponding subcarriers for subscriber stations (SS) according to channel state. It also allows multiple SSs to use a number of different subcarriers to transmit their OFDM symbols at the same time. All subcarriers in a base station (BS) are divided into sub-groups known as subchannels, which are allocated to different SSs to fulfill their various bandwidth and quality of service (QoS) requirements (Park et al, 2008; Cicconetti, Lenzini & Mingozzi, 2006; Iera, Molinaro, Pizzi & Calabria, 2007). In addition, adaptive modulation and coding (AMC) technology allows OFDMA PHY to facilitate data transmission in a high mobility environment, and makes wireless resources fully exploitable.

IEEE 802.16e now consists of five QoS scheduling types for uplink transmissions: Unsolicited Grant Service (UGS), Real-Time Polling Service (rtPS), Extended Real-Time Polling Service (ertPS), Non-Real-Time Polling Service (NrtPS), and Best Effort (BE). However, even though many packet scheduling algorithms for WiMAX and OFDMA-based networks have been proposed and tested, there are no clearly defined scheduling types for downlink transmissions or

* E-mail: rikki@asia.edu.tw
descriptions for efficiently scheduling packets in BS while taking into account their QoS (Iera et al., 2007; Sayenko et al., 2008; Nguyen & Han, 2006; Ryu et al., 2005; Shimizu et al., 2007; Kim & Kang, 2005). The five scheduling types mentioned above can be divided into two traffic flow groups for downlink transmissions: real-time (RT) and non-real time (NRT). In this paper I will propose an adaptive time-utility function (A-TUF) packet scheduling scheme to improve the QoS of RT traffic and whole system throughput. In addition to the delay and loss rate improvement for RT traffic, the proposed scheme maintains the throughput performance of NRT traffic performed by conventional TUF-based algorithms, while also enhancing system bandwidth utilization. For a base station, time-utility function (TUF) is a function that describes the delay time urgency of a head-of-line (HOL) packet. A-TUF scheme dynamically adjusts the length of a marginal scheduling time interval (MSTI) of RT’s TUF according to the queue states of NRT traffic.

The remainder of this paper is organized as follows: conventional time-utility function based scheduling algorithms are presented in Section II and the proposed scheme is described in Section III. Simulation results are offered in Section IV, and a conclusion is given in Section V.

2. CONVENTIONAL TUF BASED SCHEDULING ALGORITHMS

In OFDMA-based WiMAX network, BS should schedule packets for a limited number of SSs. A BS consists of three components: a packet classifier, a buffer management block (BMB) and a packet scheduler. The packet classifier receives the arrival packets and distributes them to the corresponding user’s sub-BMB according to the user identification number (User ID) and QoS requirement. Each sub-BMB is comprised of several different types of data buffers, with subsidiary QoS statistics (e.g., delay deadline and head-of-line [HOL] packet arrival time) being recorded for each buffer. The packet scheduler performs its task for each user according to scheduling priorities, user channel status information and QoS statistics recorded in each sub-BMB.

Many packet scheduling algorithms for the WiMAX system have been widely studied (Iera et al., 2007; Sayenko et al., 2008; Nguyen & Han, 2006; Ryu et al., 2005; Ryu et al., 2005; Shimizu et al., 2007; Kim & Kang, 2005). The urgency and efficiency based wireless (UEPS) packet scheduling algorithm (Ryu et al., 2005; Ryu et al., 2005) is considered an effective downlink scheduling algorithm for satisfying the QoS requirements of RT traffic and for maximizing NRT traffic throughput. Scheduling RT and NRT data packets for a specific user at the same time, the UEPS algorithm meets the RT QoS requirement and has higher NRT traffic throughput than the traditional Proportional Fairness (PF) and Modified Largest Weighted Delay First (M-LWDF) scheduling algorithms. UEPS uses a time-utility function (TUF) to express the time urgency values of BS data packets. TUF describes the delay time urgency of HOL packets in each sub-BMB buffer.
Figure 1. Time-utility function for NRT traffic in UEPS algorithm (Ryu et al., 2005).

Figure 2. Time-utility functions for RT, traffic in TUF-based scheduling algorithms.

Figure 1 shows the Time-utility function for NRT traffic in the UEPS algorithm. The TUF for NRT traffic of type $i$ (NRT$_i$) can be expressed as:

$$f_{\text{NRT}_i}(t) = 1 - e^{a_i t}/e^{D_i}, \quad 0 \leq t \leq D_i,$$

where $a_i$ is an arbitrary parameter of a truncated exponential function, and $D_i$ denotes the delay deadline of NRT$_i$ packets.

Function $f_i$ in Figure 2 is a TUF for type $i$ RT traffic (RT$_i$), where $d_i$ is the HOL-packet delay deadline, $j_i$ is a delay jitter, and $f_i(t)$ denotes the HOL packet TUF for traffic type $i$ at waiting time $t$. The TUF value change unit is $\left|f'_i(t)\right|$. RT$_i$ packets can be transmitted during the $[d_i - j_i, d_i]$ time interval, known as the marginal scheduling time interval (MSTI). Other NRT packets can be transmitted in the $[0, d_i - j_i]$ time interval. Therefore, RT packets will have higher priority to be transmitted if their transmission time is more urgent. When RT transmission time is not an emergency many NRT packets will be transmitted. Thus, RT QoS can be met while the NRT throughput performance can be substantially promoted.
The TUF for RT, traffic (e.g., the curve of \( f_i \) in Figure 2) in the UEPS algorithm can be written as:

\[
f_{RT} (t) = \frac{e^{a(t-c)}}{1 + e^{a(t-c)}}, \quad d_i - j_i \leq t \leq d_i,
\]

where \( a \) and \( c \) represent the slope and location of the TUF inflection point, respectively. The unit of RT, traffic utility change at inflection point \((t = c)\) is \( f_{RT} (t = c) = a \), which is applied to the RT, packet urgency factor during MSTI.

For example, the relationship between the urgency factors of all five QoS types for WiMAX uplink traffic (assuming that all SS channel states in a BS are the same) can be expressed as:

\[
|f_{RT-UGS}(t)| > |f_{RT-nOoPS}(t)| > |f_{RT-oOoPS}(t)| > |f_{NRT-oOoPS}(t)| > |f_{NRT-BE}(t)|.
\]

According to the UEPS algorithm, the BS cannot transmit RT packets earlier even if NRT packet queues are empty, since the BS must wait until the RT traffic TUF moves into the MSTI. This results in some RT packets having longer delays and occasional failures and losses due to not meeting delay deadlines. The second problem is that RT packets are still unable to be transmitted even if the NRT packet queue length is equal to zero, which may result in low bandwidth utilization. Packet scheduling algorithms should be concerned about these two issues.

Shimizu et al. (2007) therefore proposed a compromise approach involving a slight TUF modification for RT traffic. They modified the TUF for RT traffic slightly. In brief, when the NRT queue length is greater than zero, the TUF for RT, traffic is identical to that set by the UEPS (e.g., the curve of \( f_i \) in Figure 2); when NRT queues are empty, the TUF for RT, traffic equals \( f_2 \) in Figure 2, which can be written as:

\[
f_{RT} (t) = \frac{e^{a(t-c)}}{1 + e^{a(t-c)}}, \quad 0 \leq t \leq d_i.
\]

In this situation, MSTI obviously changes into the maximum allowable delay. In a special case (NRT queue length = 0), Shimizu's algorithm can schedule the RT packets earlier so that RT packet delay can be reduced (RT QoS is improved) and system throughput can also be increased.

### 3. PROPOSED SCHEDULING SCHEME

However, Shimizu et al.'s (2007) method still lacks sufficient flexibility. First, it assumes that the number of RT packets in a BS is larger than the number of NRT packets. They may be right over the long term but wrong in the short term. Second, their method differs little from the UEPS algorithm in that the MSTI length of TUF for RT traffic is set to the maximum only when all NRT queues are empty. I
therefore propose an Adaptive TUF (A-TUF) scheduling scheme to improve RT packet delay, RT packet loss rate and whole system throughput.

Figure 3. Adaptive time-utility function for RT\textsubscript{i} traffic in the proposed A-TUF scheduling scheme.

A-TUF dynamically adjusts the MSTI length of TUF for RT traffic according to the ratio of average queue length (L\textsubscript{NRT}) to average queue size (Q\textsubscript{NRT}) of all NRT buffers. As shown in Figure 3, the TUF \( f_\theta \) for RT\textsubscript{i} traffic can be defined as:

\[
    f_{RT_i}(t) = \frac{e^{-\theta_i(t-t_c)}}{1 + e^{-\theta_i(t-t_c)}} , \quad \theta_i \leq t \leq d_i ,
\]

with \( \theta_i \) expressed as:

\[
    \theta_i = \left( \frac{d_i - j_i}{m} \right) \leq \sum_{k=1}^{m} \frac{L_{NRT_k}}{Q_{NRT_k}}
\]

where \( m \) denotes the number of scheduled NRT traffic types in a BS.

\[
    \text{MSTI}_{\theta_i} = d_i - \theta_i , \quad 0 \leq \theta_i \leq (d_i - j_i).
\]

If the number of NRT packets in a BS is equal to zero, then \( \sum L_{NRT_k} = 0 \). That is, \( \theta_i = 0 \) and \( \text{MSTI}_{\theta_i} = \text{MSTI} \). The flexibility of the MSTI length is changed to the maximum allowable delay, i.e., \( \text{MSTI}_{\max} = d_i \). The dynamic variation in MSTI length for RT\textsubscript{i} traffic can be written as:

\[
    j_i = \text{MSTI}_{\min} \leq \text{MSTI}_{\theta_i} \leq \text{MSTI}_{\max} = d_i.
\]

Figure 4 shows the maximum and minimum values of MSTI length for RT\textsubscript{i} traffic.
Furthermore, how to allocate OFDMA subchannels to SSs properly is very important in the subsequent subchannel allocation stage, because the frequency fading characteristics of the subchannels are different. According to the proposed A-TUF scheme for deciding transmission priority, a BS allocates the smallest frequency fading subchannel to the first priority HOL packet and the next smallest to the second priority HOL packet. It then allocates all remaining subchannels to other packets according to the same policy until all channel frames are used.

**Figure 4.** Maximum and minimum MSTI of dynamic time-utility function for RT_i traffic in the proposed A-TUF scheduling scheme.

**Figure 5.** Comparison of VoIP average delay between A-TUF and the other two schemes.
Figure 6. Comparison of Video average delay between A-TUF and the other two schemes.

Table 1. Simulation parameters

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total number of OFDM subcarriers</td>
<td>1536</td>
</tr>
<tr>
<td>Number of subchannels</td>
<td>12</td>
</tr>
<tr>
<td>Number of subcarriers per channel</td>
<td>128</td>
</tr>
<tr>
<td>OFDM frame period</td>
<td>20ms</td>
</tr>
<tr>
<td>Slot period</td>
<td>1ms</td>
</tr>
<tr>
<td>Channel estimation</td>
<td>Perfect</td>
</tr>
<tr>
<td>Modulation and coding</td>
<td>AMC 16-QAM, Rate 2/3</td>
</tr>
<tr>
<td>ON period of VoIP audio</td>
<td>Exponential (mean=1.0sec)</td>
</tr>
<tr>
<td>OFF period of VoIP audio</td>
<td>Exponential (mean=1.5sec)</td>
</tr>
<tr>
<td>Inter-arrival time between Video frames</td>
<td>100ms</td>
</tr>
<tr>
<td>Number of packets per Video frame</td>
<td>8</td>
</tr>
<tr>
<td>Video packet size</td>
<td>Geometric (mean=50bytes)</td>
</tr>
<tr>
<td>Inter-arrival time between FTP packets</td>
<td>Exponential (mean=40sec)</td>
</tr>
<tr>
<td>Buffer size for FTP traffic</td>
<td>500Kbytes</td>
</tr>
<tr>
<td>FTP packet size</td>
<td>Geometric (mean=20Kbytes)</td>
</tr>
<tr>
<td>Inter-arrival time between Email packets</td>
<td>Exponential (mean=500sec)</td>
</tr>
<tr>
<td>Buffer size for Email traffic</td>
<td>1Mbytes</td>
</tr>
<tr>
<td>Email packet size</td>
<td>Geometric (mean=200Kbytes)</td>
</tr>
<tr>
<td>Simulation software</td>
<td>C programming</td>
</tr>
</tbody>
</table>
4. SIMULATION RESULTS

To compare the performance of the proposed A-TUF scheme to that of UEPS and Shimizu et al.’s algorithm, I designed a simulation using four traffic types: RT-UGS (VoIP), RT-rtPS (Video), NRT-nrtPS (FTP), and NRT-BE (Email), assuming that VoIP allows a maximum packet delay in a BS of 40 ms and Video a maximum of 150 ms. The simulation uses a Markov ON-OFF model to generate regular VoIP voice packets at a fixed size, and a Video generator to produce variable-sized video packets. Assume the downlink is based on a MIMO-OFDM air interface, with the average coding rate 2/3, 16-QAM modulation, and spatial multiplexing mode, as shown in (Dubuc, Starks, Creasy & Hou, 2004). Given a carrier with a 20MHz bandwidth, the system divides the carrier to 1536 subcarriers (12 subchannels). The simulation was performed with various values of system load ranging from 0.455 to 1.307. The inter-arrival time for different kinds of traffic was distributed exponentially with different mean values, while the packet size has a geometric distribution. Other simulation parameters are listed in Table 1.

Results from a comparison of the VoIP average delay indicate strong similarities between the proposed A-TUF scheme, UEPS and Shimizu et al.'s algorithm (Figure 5). The main characteristics of VoIP packets are fixed size and high priority but, since they don’t have large bandwidth requirements, their QoS requirements are easily met. In terms of Video average delay the results indicate better performance for A-TUF compared to the other two algorithms (Figure 6), perhaps due to its ability to compute and adjust MSTI length to allow for earlier Video packet transmission. According to these combined results, A-TUF outperformed the other two algorithms, especially in cases of light system load.

Moreover, RT packets still waiting for long periods beyond delay deadlines are generally considered as failed and therefore dropped/lost. Results from a comparison of average Video traffic loss rate among the three algorithms are shown in Figure 7; they indicate a slightly lower loss rate for the proposed A-TUF, again due to its ability to adjust the MSTI. A-TUF is capable of buffering rapid traffic bursts and scheduling RT packets earlier, resulting in reduced RT traffic loss.

Results for FTP and Email throughput performance comparisons are not shown due to the insignificant differences noted between the three approaches. However, A-TUF is capable of maintaining stable throughput even if RT traffic gets a better QoS, since the higher the number of NRT packets in a queue, the greater the $\theta$ for RT traffic—that is, MSTI shrinks and gets closer to MSTI$_{min}$.

5. CONCLUSION

In summary, the proposed A-TUF scheme introduces lower RT packet delay and loss while still maintaining the same NRT throughput performance as conventional TUF-based algorithms. At the same time it enhances system bandwidth utilization and maintains an acceptable degree of fairness between RT and NRT traffic.
We are now continuing to study how to integrate other packet scheduling techniques, such as Adaptive - Delay Threshold-based Priority Queuing (Kim & Kang, 2005; Kang, Kim & Kim, 2008), to achieve the best usage of wireless bandwidth resources, to meet SS’s QoS highly, to improve the system throughput more, and to provide OFDMA-based WiMAX systems the best scheduling implementation.

![Figure 7. Comparison of Video average loss rate between A-TUF and the other two schemes.](image)

**REFERENCES**


Amendment for Physical and Medium Access Control Layers for Combined Fixed and Mobile Operation in Licensed Bands.


**Jui-Chi Chen** received his B. S. (1993) and M. S. (1995) in Computer Science and Information Engineering from National Chao Tung University and Ph. D. (2006) in Computer Science from National Chung Hsing University. He is currently an Assistant Professor in the Department of Computer Science and Information Engineering at Asia University, Taiwan. His research interests include wireless communications and computer networks.