Measurement and Analysis of One-Way Delays over IEEE 802.16e/WiBro Network

Dongmyoung Kim*, Hua Cai†, and Sunghyun Choi‡

*School of Electrical Engineering and INMC, Seoul National University, Seoul, Korea
†Samsung SDS Co. (China), Beijing, China
‡Email: dmkim@mwnl.snu.ac.kr, hu1.cai@samsung.com, schoi@snu.ac.kr

Abstract—One-way delay is an important metric to evaluate the performance of communication networks. In this paper, we analyze the one-way delay performance of the commercial WiBro networks based on the traffic measurement. In measurements, much large delay is observed in the uplink than downlink, and we find that such asymmetry due mainly to the impact of the uplink bandwidth request mechanism. Due to this large uplink delay, the performance of commercial IEEE 802.16e service can be severely degraded. We also discuss some possible solutions to ameliorate this problem.

I. INTRODUCTION

WiBro (Wireless Broadband) is a Korean version of Mobile WiMAX/IEEE 802.16e system, which is designed to provide broadband data services in a mobile environment [1]. WiBro operates at 2.3 GHz bands with a 10 MHz channel bandwidth, and it employs orthogonal frequency division multiple access (OFDMA) and time division duplexing (TDD) schemes. In 2006, WiBro systems were deployed in Seoul, Korea, and the world first commercial service based on Mobile WiMAX/IEEE 802.16e commenced. Detailed information about Mobile WiMAX and WiBro is provided in [2].

The performance of IEEE 802.16e system has been analyzed by many researchers based on either computer simulations or numerical analysis [3,4]. However, both simulations and analysis have inherent limitations since they are developed from numerical models and the assumptions for realities. In this context, the performance evaluation based on measurement is very essential since we can obtain the realistic view on the working system. Recently, there have been some efforts to evaluate the performance of WiBro system based on measurements in live operational networks. In our previous work, we analyzed the throughput and delay performance of WiBro system in various environments [5]. Similarly, QoS of VoIP applications is evaluate based on the measured throughput, delay and packet losses in [6]. It was shown that the round trip time (RTT) delay performance of WiBro system is poor compared with the conventional data networks, e.g., Ethernet or Wi-Fi networks, though the throughput performance is quite good. However, as a performance metric, RTT has a limitation. For example, RTT does not have any meaning with streaming applications, such as VoIP or video streaming, which produce traffic flows in one direction. For such applications, we need to understand one-way delays in both directions. Furthermore, one-way delay results help us to analyze the cause of long delivery delay in WiBro networks.

Motivated by this, we present measurement results for one-way delay of a commercial WiBro system and analyze them. From the measurement results, it is found that uplink direction involves long delays due mainly to contention-based bandwidth request operations.

The rest of the paper is organized as follows. We first describe measurement environments in Section II. Then, the measurement results are given and analyzed in Section III. In Section IV, some possible solutions to reduce uplink one-way delay are introduced. Finally, we conclude this paper in Section V.

II. MEASUREMENT SET-UP

We have conducted our measurements in a commercial WiBro networks in the campus of Seoul National University (SNU), Seoul, Korea. A single base station (BS) is located about 200 meters away from our measurement spot, i.e., the 4th floor at building 132 in SNU. For the WiBro host machine, we use a labtop and connect it to the with WiBro access network via a Samsung SPHH1100 PCMCIA type WiBro card.

The experiments to measure the RTT values are conducted in the environment shown in Fig. 1. For the RTT measurement, a time stamp is inserted in a small packet and the packet is sent over the WiBro link from a laptop connected to the WiBro network to a desktop connected to the wired Internet. Then, the desktop sends back an answering message. Upon receiving the message, the laptop determines the RTT by comparing the time stamp in the received packet with the current local time. For the comparison purpose, the same experiments are conducted for a wireless LAN (WLAN), i.e., Wi-Fi, system in the similar measurement environment.

Measuring one-way delays is more challenging than measuring RTT. Note that in the one-way delay measurement, the elapsed time should be calculated in the receiver while the time stamp is inserted in the different device, i.e., sender. Therefore, the receiver should be synchronized with the sender in order to correctly determine the one-way delay based on the time stamps. The network based the synchronization, e.g.,
Network Time Protocol (NTP), is usually used for synchronization among network nodes. For example, IxChariot [7], a commercial traffic generation/measurement tool, uses NTP for synchronization. However, we have found that the measured one-way delay results are very unreliable when the synchronization is done using NTP due to the large dynamics in the wireless access system. In [8], the authors develop the QoS measurement tool, which employs a Global Positioning System (GPS)-based synchronization to overcome this problem.

In our measurements, we synchronize the receiver with the sender in another novel manner. We install both Ethernet card and WiBro card to a single mobile station (MS). Then, the downlink packets are transmitted via Ethernet card, and destined to the WiBro card of the MS after going through the Internet. In this measurement, a perfect synchronization can be achieved since the sender and the receiver are actually the same. We can also obtain the uplink one-way delay using the similar manner. The environment for the one-way delay measurement is also illustrated in Fig. 1.

### III. Measurement Results and Analysis

#### A. Baseline Delay Measurement

We have conducted experiments to measure both RTT and one-way delay in the environments described in Section II. Measurement packets are transmitted at every one second without any other background traffic in the same device, and the procedure is repeated 500 times. In Table I, the averaged results are presented. In the results, we observe that the RTT of WiBro is very large compared with that of WLAN system. Similar results are reported in [5].

In order to analyze the detailed delay statistics, we also measured one-way delays for both uplink and downlink. As a result, much larger delay is observed in the uplink direction than the downlink direction as shown in Table I. Here, the end-to-end one-way delay is composed of sender-related, radio access, radio network, Internet and receiver-related delays. Among the delay components, we expect that the radio access delay is the major and the most dynamic component while the other delay components are rather minor and can be roughly estimated as stable values provided that the network is not congested. The radio access delay is closely related to the radio access protocol of the WiBro system. Therefore, the uplink and downlink radio access delays can be different because of the differences between the uplink and downlink mechanisms.

One major unique procedure in uplink packet transmission is the uplink bandwidth request. In the WiBro system, the resource allocation is done at the BS in a centralized manner. When a downlink packet arrives at the queue of a BS, the BS becomes instantly aware of it. Hence, the resources for the packet can be easily allocated based on the packet size and the transmission rate to the destination over the air. On the other hand, when an uplink packet is generated at an MS, the BS cannot instantly allocate resources for the packet because BS is not aware of the existence of the newly generated packet. Therefore, additional procedures are required to inform the existence as well as the required resources for the uplink transmissions.

In IEEE 802.16e/WiBro system, five methods are defined for uplink bandwidth request, namely, unicast/multicast polling, piggy-backing, bandwidth stealing, contention-based bandwidth request, and the usage of Poll-Me (PM) bit in the Unsolicited Grant Service (UGS) connections. All the five methods impose some delay before allocating the uplink resources. Among the methods, the contention-based bandwidth request and piggy-backing are generally used for the Best Effort (BE) connections, and all the connections are served as BE connections in the current commercial WiBro system, deployed in Seoul.

The procedures for the two bandwidth request schemes used for BE connections, i.e., contention-based bandwidth request and piggy-backing, are illustrated in Fig. 2. In the contention based method, an MS takes the following four-step request-response procedure as shown in Fig. 2(a). The MS choose a random bandwidth request ranging (BR_RNG) code, which is modulated into the contention-based ranging sub-channel, composed of a number of Orthogonal Frequency Division Multiplexing (OFDM) sub-carriers, and transmitted to the BS. Note that this code transmission is contention-based, so that a code transmission might fail due to a collision with other codes. If the BS detects the BR_RNG code successfully,
the BS sends a message in the UL-MAP\(^1\), which indicates the transmission region for the bandwidth request message. The MS sends a stand-alone bandwidth request (BR) MAC header which specifies the required amount of resources, and the required resources are allocated by the BS through a subsequent UL-MAP. Finally, the MS transmits its uplink packet by using the allocated resources. On the other hand, the MS can send an uplink bandwidth request using the piggybacking method as shown in Fig 2(b). The bandwidth request message for the newly-generated packets can be piggy-backed in the sub-header of ongoing MAC frame in the MS. Since the sub-header specifies the required amount of resources, the BS can allocate the resources for uplink packet in two steps.

Compared with the contention-based bandwidth request, piggyback imposes smaller delay because it requires only two message exchanges before the resource allocation and it does not require contending procedures to get the bandwidth request opportunity. Though the piggy-backing is efficient, piggy-backing can be utilized only in limited situations where other uplink packets are scheduled and transmitted. In the previous measurements, piggy-backing cannot be used because only the measurement packets are transmitted every one second without any background traffic. Therefore, we can expect that the difference between uplink and downlink delays mostly comes from the uplink bandwidth request delay.

B. Delay Measurement with Background Traffic

To investigate the impact of contention-based bandwidth request, we conducted measurements in the environments where a background uplink traffic flow exists. Because the chance to utilize the piggy-backing is related to the existence of the already allocated resources, we control it by making background traffic. Here, we generate uplink packets periodically with a fixed interval as background traffic. Then, we generate measurement packets with the interval of 2 seconds, and measure one-way delay. We conducted the same experiments by increasing the transmission interval of the background traffic packets from 2 msec to 150 msec.

The results are shown in Fig. 3. We find that the one-way delay decreases as the transmission interval of the background traffic packets decreases. At the first glance, it is quite counterintuitive since heavier background traffic results in shorter delay. In reality, the more frequently the background packets are transmitted, the more opportunities for piggy-backing bandwidth request the one-way delay-measured flow will have, thus decreasing the uplink one-way delay. This should be true as long as the background traffic does not overload the network. When the transmission interval of the background traffic is 2 msec, the measured uplink one-way delay is almost the same as the downlink one-way delay. From the results, we conclude that the bandwidth request delay of WiBro network is one of the major delay components especially for stand-alone BE connections. Therefore, more research and development on the efficient usage of the bandwidth request is required.

C. Detailed Analysis for the Bandwidth Request Schemes

In this section, we estimate the amount of bandwidth request delay in contention-based bandwidth request based on a simple analysis. Based on the procedures of bandwidth request mechanisms explained in Section III-A, we can divide the uplink one-way delay into the following five delay components when the contention-based bandwidth request is utilized.

\(^1\)In the WiBro, a UL-MAP appears in every MAC frame, where MAC frames repeat every 5 msec.
1) Elapsed time between data arrival and random access code generation at an MS (defined as $T_0$).
2) Elapsed time between random access code generation at the MS and successful reception at a BS (defined as $T_1$).
3) Elapsed time between random access code reception and UL-MAP transmission for bandwidth request at the BS (defined as $T_2$).
4) Elapsed time between UL-MAP reception and bandwidth request header transmission (defined as $T_4$).
5) Elapsed time for all the other procedures (defined as $T_8$).

$T_1$ is large in some environments where code retransmissions occur because it includes random access delay. However, in the lightly loaded systems, $T_1$ is not very large. $T_2$ can also be large due to the complexity for decoding the random access code at the BS. It is known that several tens milliseconds of delay is required in many current devices. $T_3$ is configured by a BS and notified to MSs in the WiBro system. $T_3$ is generally set to several frames because it is difficult to decode UL-MAP message instantly and use the allocated resource in the same frame. Therefore, several tens milliseconds of delay can be imposed for $T_3$ considering the frame length of WiBro, i.e., 5 msec.

On the other hand, only the following three delay components exist for piggy-backing bandwidth request.

1) Elapsed time between data arrival and generation of a bandwidth request sub-header for piggy-backing (defined as $T_0'$).
2) Elapsed time between sub-header generation and piggy-backed packet transmission (defined as $T_3'$).
3) Elapsed time for all the other procedures ($T_8'$).

We refer to the sum of $T_1$ and $T_2$ and $T_3$ in contention-based bandwidth request as $T_c$, and $T_1'$ in piggy-backing is re-defined as $T_p$. We expect that difference between $T_c$ and $T_p$ causes the one-way delay differences. $T_4$ is excluded in $T_c$ and $T_p$ because it represents the common delay component after getting chance to transmit the bandwidth request, and we ignore $T_0$ and $T_0'$ because they are small values and the two values are almost the same.

When the background flow exists, both bandwidth request schemes can be utilized. If $T_c$ is smaller than $T_p$, the contention-based bandwidth request is used, while piggy-backing is used with the smaller $T_p$. Therefore, the difference between the uplink one-way delays with and without background traffic is given as $d_{\text{analysis}} = T_c - \min(T_c, T_p)$. Here, $T_p$ is a function of $r$ which is defined as the interval of the background packets. Because the measurement packet generation and background packet generation are independent, we can assume that the probability distribution function (probability density function) of $T_p$, i.e., $f_{T_p}$, is the uniform distribution over $[0, r]$. Then, $d_{\text{analysis}}$ can be shown as follows with given $r$:

$$d_{\text{analysis}} (r) = \int_0^{\infty} (T_c - \min(T_c, t)) f_{T_p} (t) \, dt$$

$$= \int_0^{\min(T_c, r)} \frac{1}{r} (T_c - t) \, dt$$

In the previous section, we have shown some experimental results with various background packet intervals, i.e., $r$. We refer to the difference between the uplink one-way delays with and without background traffic measured in the experiments as $d_{\exp}$. In 1, $d_{\text{analysis}}$ is expressed as a function of $r$ and $T_c$. Because $r$ used in each experiment is known, we can estimate the unknown parameter $T_c$ by comparing $d_{\text{analysis}}$ and $d_{\exp}$. Fig. 4 shows $d_{\exp}$ and $d_{\text{analysis}}$ calculated with various assumptions for $T_c$. From the results, we can conclude that $T_c$ value is about $45 \sim 50$ msec, which is a very large value. As explained before, all the three components in $T_c$, i.e., $T_1$ and $T_2$ and $T_3$, can impose the significant delay in the current standard/implementation. Therefore, we need to find some ways to reduce this bandwidth request delay in the contention-based bandwidth request.

IV. POSSIBLE SOLUTIONS

In this section, some possible methods to reduce the large uplink bandwidth request delay are discussed. Note that proposing a novel scheme is out of scope for this work. Hence, we just categorize the possible methods and introduce some existing solutions for this problem. The methods to reduce the uplink bandwidth request delay can be categorized into the following four approaches:

1) Improving the performance of devices.
2) Establishing proper types of connections defined in IEEE 802.16e standard.
3) Developing novel algorithms for bandwidth request and allocation.
4) Updating protocols in the standard.

Firstly, improved hardware performance can reduce the uplink bandwidth request delay. For example, the $T_2$ and $T_3$ defined in the previous section are closely related to the hardware performance of BS and MS, respectively. Therefore,
some amount of delay can be reduced in the future by using the better devices.

Secondly, the problem for the real-time applications can be reduced by defining other types of connections such as UGS, rtPS, ertPS, nrtPS which generally utilize the other kinds of bandwidth request methods defined in [1]. In the other types of connections, uplink bandwidth delay can be reduced based on the proactive bandwidth reservation. However, problems still exist because of following reasons: 1) as shown in [5], the best-effort type of flows can also be influenced by delay performance if the best-effort flows utilize TCP; 2) only the BE connection might be used in the early stage of system deployment, like the current WiBro system in Seoul, Korea.

Thirdly, some resource allocation algorithms can be used to reduce the uplink bandwidth request delay without changing the current IEEE 802.16e/WiBro standard. For example, some bandwidth allocation schemes have been proposed to improve the performance of downlink TCP traffic by reducing the bandwidth request delay for transmitting TCP ACK in uplink direction [9, 10]. Though the uplink TCP ACK traffic is correlated to the downlink TCP data traffic, the bandwidth allocation for TCP ACK is basically decoupled from that for TCP data. Motivated by this, the bandwidth request mechanism in [9] reduces the uplink bandwidth request delay by allocating resource for bandwidth request MAC header or TCP ACK message in the proactive manner. However, such a proactive scheme cannot determine the accurate amount of the required resource. To overcome these drawbacks, the authors in [10] propose a hybrid approach which basically employs piggy-backing and utilizes proactive scheme only if piggy-backing is not available. In [10], the scheduler in the BS checks the uplink request queue of the corresponding MS on serving a downlink frame. If the request queue is empty, the scheduler puts a new bandwidth request in a proactive manner on behalf of MS. Then, the allocated bandwidth can be used for uplink TCP ACK, and if there remains backlogged traffic in MS’s transmission queue, the MS requests bandwidth using the piggy-backing.

Fourthly, the protocols defined in the WiBro standard can be updated to ameliorate this problem. A two-way contention-based bandwidth request scheme is proposed in [11] for the on-going IEEE 802.16m standard. IEEE 802.16m standard has been discussed in IEEE 802.16 Task Group m (TGm) to make a descending standard of IEEE 802.16e [12]. The current contention-based bandwidth request in IEEE 802.16e system requires four steps to allocate the resources because only the bandwidth request opportunity is given to an MS after a BS receives the first-transmitted random access code from the MS. On the other hand, in [11], the amount of required bandwidth is embedded in the first random access code by modifying some parts of the random access code. Then, the bandwidth request delay is reduced because the bandwidth allocation is done in two steps. However, some problems still exist in [11]. Ability for the contention-resolution is degraded by destructing the random access code to embed the bandwidth request. Furthermore, some amount of uplink resources can be wasted because the embedded information should be coarsely quantized.

V. Conclusion

In this paper, we analyze the one-way delay performance of WiBro system via measurements. It is measured that the uplink one-way delay is much larger than the downlink one-way delay. We show that this asymmetry mostly comes from the large delay incurred by the contention-based bandwidth request mechanism. Because this large uplink delay can degrade the system performance, we also introduce some possible methods to reduce the delay.

Acknowledgments

This work was in part supported by the IT R&D program of MKE/IITA (2008-F-007-02, Intelligent Wireless Communication Systems in 3 Dimensional Environment) and the Information Technology Research Center support program of MKE/IITA (IITA-2009-C1090-0902-0006).

References

[12] IEEE 802.16m, IEEE 802.16 Task Group m (TGm), Official IEEE 802.16m Website: http://www.ieee.org/16/gtm/.