An Adaptive Bandwidth Request Scheme for QoS Support in WiMAX Polling Services

Cheng-Yueh Liu and Yaw-Chung Chen
Department of Computer Science
National Chiao Tung University, Hsinchu, Taiwan
ycchen@cs.nctu.edu.tw

***Abstract*---**IEEE 802.16, also called WiMAX was developed to accommodate large coverage and high bandwidth last-mile Internet access. In WiMAX, provisioning of QoS is still an important issue. Before transmitting packets, subscriber stations (SS) must send a bandwidth request message to the base station (BS). If available bandwidth is not sufficient to use unicast polling for each SS to send bandwidth request, the BS will form a group of stations that utilize multicast and broadcast polling to contend the slots for transmitting bandwidth request. Since extended real-time polling service (ertPS) proposed in 802.16e has strict delay time requirement, the lower the delay, the better the QoS. In order to reducing the delay caused by collisions of request packets, we proposed a scheme that utilizes contention-free period to allocate slots for bandwidth request. Both mathematic analysis and simulation results show that our scheme features lower delay and better QoS performance for ertPS flows as the number of stations increases.

***Keywords*---**IEEE 802.16, WiMAX, QoS, bandwidth request, contention period

1. Introduction

The Worldwide Interoperability for Microwave Access (WiMAX) based on the IEEE 802.16-2004 [1] Air Interface Standard rapidly advances as a main stream technology in fixed broadband wireless metropolitan area networks. In 2005 the IEEE 802.16e amendment adds features and attributes to the 802.16 standard [2] for supporting mobility. Beyond the air interface, the WiMAX Forum is defining the network architecture necessary for implementing an end-to-end mobile WiMAX network [3]. It is expected that WiMAX technology should be incorporated in notebook computers and PDAs by 2007, allowing for urban areas and cities to connect with each other for portable outdoor broadband wireless access.

QoS (Quality-of-Service) is an important issue in wireless network especially for voice and video flows. IEEE 802.16 has already specified complete construction, including how they achieve QoS requirement. There are five types of scheduling services specified for different traffic models, i.e. Unsolicited Grant Service (UGS), real-time Polling Service (rtPS), non-real-time Polling Service (nrtPS), Best Effort (BE), and extended real-time Polling Service (ertPS) published in IEEE 802.16e. Among these five service flows, the QoS support in UGS, rtPS, and ertPS is necessary.

In 802.16 PMP (point to multipoint) mode, a subscriber station must send a bandwidth request to base station before it transmits the uplink data. When the available bandwidth isn't sufficient for all stations to make bandwidth request separately, the service flows belonging to ertPS, nrtPS, and BE have to do contention to send their requests. There is certain collision probability during contention period and that would cause the request delay for the ertPS flows. This situation may result in the station unable to get proper bandwidth so that its QoS will be degraded. The above collision problem is addressed in this study.

The rest of this paper is organized as follows. Section 2 describes the background for this study. Section 3 discusses the problem in 802.16 and our proposed scheme in detail. The numerical analysis and simulation evaluation are presented in Section 4. Conclusions and future works are stated in Section 5.

2. Background

Five services are supported in IEEE 802.16e, by specifying a scheduling service with associated QoS parameters, the BS scheduler can anticipate the throughput and provide polls and/or grants at the appropriate times.

2.1 Bandwidth allocation and request mechanisms

Whenever an SS needs to ask for bandwidth for a connection, it sends a message containing the immediate requirements to the BS. QoS for the connection was established at connection establishment and is looked up by the BS.

Requests
The Bandwidth Request message may be transmitted during uplink bandwidth allocation, except during initial ranging interval. Bandwidth Requests may be incremental or aggregate, when the BS receives a Bandwidth Request from any SS, it should perform admission control to check whether the request is granted or not.

Polling
According to the bandwidth availability, there are three polling types, unicast, multicast, and broadcast. Unicast polling could only be used if the bandwidth is sufficient for polling all SSs individually. Otherwise some SSs may

1545-0678/08 $25.00 © 2008 IEEE
DOI 10.1109/ICDCS.Workshops.2008.77

be polled by multicast polling or broadcast polling, and contention resolution algorithm is used in multicast or broadcast polling to reduce the collision probability.

2.2 MAC support of PHY

There are two major duplex techniques: Frequency Division Duplex (FDD), and Time Division Duplex (TDD). TDD is the preferred duplex mode for the following reasons: TDD allows adjustment of the downlink and uplink ratio to efficiently support asymmetric two-way traffic; TDD only requires a single channel for both downlink and uplink. It also can provide greater flexibility for adaptation to varying global spectrum allocations; TDD system designed by factories is less complex and therefore less expensive.

2.3 Related Works

Most of 802.16e scheduling algorithms were focusing on QoS issues especially for UGS, rtPS, and ertPS. In [4], the author proposed a simple and efficient algorithm which calculates each connection allocation slots and controls order of slots to reduce the jitter. In [5], the proposed scheme adaptively allocates bandwidth in order to control queue occupancy at a target level so that the requirements for delay and PDU dropping probability can be met. In [6], authors suggest a system model that uses voice activity detector and silence detector to check whether the state is on or off, and relies on system to calculate the steady-state probability.

Several adaptive schemes were proposed to dynamically allocate resource allocations. The scheduling problem is to allocate time slots on a subset of sub-carriers to confirm client's demands and maximize system throughput. In [7], it presented linear programming relaxations for resource allocation problem and provide optimal allocations for all users.

IEEE 802.16e provides ertPS for supporting silence suppression, it has some advantages that not only reduce MAC overhead and access delay of the rtPS algorithm, but also prevent the waste of uplink resources for the UGS algorithm. In [8], the authors make some analyses to verify the ertPS performance.

3. Proposed Approaches

IEEE 802.16e ertPS supports real-time service flows that generate variable size data packets on a periodic basis, such as VoIP services with silence suppression. It would be very likely for users to use handset devices in place of traditional telephones to make VoIP calls, which may utilize 802.16e ertPS service. In WiMAX system, when bandwidth isn't enough for unicast polling, ertPS must use multicast polling or broadcast polling to participate contention with nrtPS or BE. This usually causes collisions during contention period so that the bandwidth request delay could not be guaranteed, and the packets could not be transmitted immediately.

Based on the above observation, we proposed a scheme for providing better QoS in WiMAX environment. The scheme adjusts their transmitting sequence to allow ertPS connections to have higher priority. We use both mathematic and simulation approaches to evaluate the performance.

3.1 Main Scheme

Figure 1 shows the original multicast or broadcast polling in which each slot is a transmission opportunity. If collision occurs, backoff will be performed to make request again. Our goal is to reduce the request collisions for ertPS flows. The key idea of our scheme is to steal some request contention time of nrtPS and BE services, because they do not have immediate QoS demand. If there were VoIP packets to be sent through ertPS flows, we assign the first several slots to ertPS services and these assigned slots could not be used by either nrtPS or BE flows.

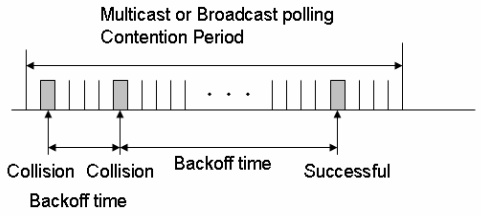


Figure 1 Slots allocation of original scheme.

We divide the original polling contention period into two parts, the contention-free (CF) period and the contention period. The heading slots are integrated into CF period which could be allocated to ertPS flows only, and the remaining slots keep the original functionality. Figure 2 illustrates the proposed slots allocation. If an ertPS flow was allocated a CF period slot, it can transmit the bandwidth request packet immediately; otherwise, it can participate the contention period with nrtPS flows and BE flows.

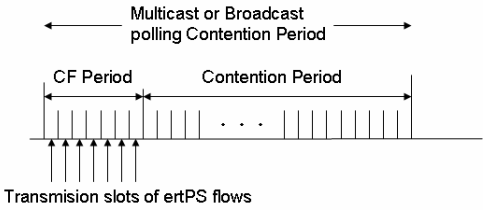


Figure 2 Slots allocation of proposed scheme.

3.2 Contention-free period definition

Each CF slot can only be used by the assigned service flow, even if there is no request to send. In general, it is not sufficient for allocating the CF slots to service flows one by one. The BS won't be aware when the SSs have bandwidth request to send, so we allocate each CF slot to ertPS flow with random selection for fairness. Although we assign some slots to let ertPS flows transmit bandwidth request first, the number of CF slots still has

its limitation in providing sufficient bandwidth for ertPS and BE flows. So we set threshold to one half of the contention period.

Figure 3 shows a slots allocation sample of our scheme. We assume that the randomly selected number is 4, so the CF slots are allocated to the ertPS flows whose serial numbers are from 4 to 8. Slots 4 and 7 indicate that stations are transmitting bandwidth requests; other slots (5, 6, and 8) are idle because the station may not be ready for transmission. Other ertPS flows without CF slots could only participate contention period with nrtPS and BE flows.

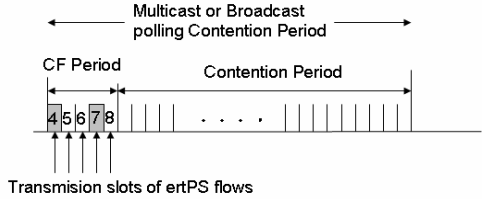


Figure 3 Slots allocation example of proposed scheme.

3.3 Mathematical analysis

We use mathematical analysis to compare the performance of our scheme with the original scheme in the standard. We define the delay as the time duration from a station is ready until its bandwidth request message has been successfully transmitted.

Numerical analysis of the original scheme

Assume C slots are allocated for contention period, the probability (P_{suc}) of successfully transmitting a bandwidth request message during the contention period can be derived as follows:

$$P_{suc} = \frac{1}{C}\left(1-\frac{1}{C}\right)^{N-1} \times C = \left(1-\frac{1}{C}\right)^{N-1} \quad (1)$$

While the collision probability (P_{col}) of transmitting a bandwidth request is:

$$P_{col} = 1 - P_{suc} = 1 - \left(1-\frac{1}{C}\right)^{N-1} \quad (2)$$

We use a specific signaling channel to implement contention period. A frame could only have one signaling channel, so the collision frequency in a signaling channel is same as that in a frame. Using geometric distribution, we can derive the probability ($P_{suc(k)}$) of transmitting a successful bandwidth request until the *k-th* frame:

$$P_{suc}(k) = P_{col}^{k-1}\left(1 - P_{col}\right) \quad (3)$$

Equation (3) shows the probability that retransmissions during the first (k-1) frames collided with other service flows and won't be successful until the *k-th* frame. We then use an expected value E to represent the average number of retransmission for a bandwidth request message:

$$E = \sum_{k=1}^{\infty} kP_{suc}(k) = \frac{1}{\left(1-\frac{1}{C}\right)^{N-1}} \quad (4)$$

We convert the expected value to an expected collision count. Beside the expected collision count, we also need to consider the back-off delay [11].

$$Backoff\_delay = \frac{Backoff\_slot}{slots\_perframe} * frame\_size \quad (5)$$

Where Equation (5) may cause some inaccuracy, because we do not know when the back-off starts. Since the difference is not too far from what we expect, the inaccuracy is acceptable. Finally, we combine the expected value and back-off delay to get real collision delay:

$$\begin{aligned} Delay &= Frame\_size \times E + Backoff\_delay \\ &= \frac{F}{\left(1-\frac{1}{C}\right)^{N-1}} + Backoff\_delay \end{aligned} \quad (6)$$

Where F is the frame size with a typical value 5 ms.

Numerical analysis of proposed scheme

The probability of successfully transmitting a bandwidth request message during a frame (or a contention period) can be derived as follows [11]:

$$P_{suc} = \frac{\sum_{i=1}^{CF}\left(1-\frac{1}{C-i}\right)^{N-i}}{CF} \quad (7)$$

Where *C* is the number of total slots, *CF* is the number of slots in contention free period, *N* is the number of total service flows (all service flows can participate contention including ertPS flows), and *i* means a variable number of service flows that do not participate the contention.

$$P_{col} = 1 - P_{suc} = 1 - \frac{\sum_{i=1}^{CF}\left(1-\frac{1}{C-i}\right)^{N-i}}{CF} \quad (8)$$

$$P_{suc}(k) = P_{col}^{k-1}\left(1 - P_{col}\right) \quad (9)$$

Equation (8) represents the probability of unsuccessful transmission of a bandwidth request. Equation (9) means that bandwidth request message would be transmitted successfully in the *k-th* frame, so the order of P_{col} must be *k-1*.

$$E = \sum_{k=1}^{\infty} kP_{suc}(k) \quad (10)$$

$$\begin{aligned} Delay &= frame\_size \times E + Backoff\_delay = \\ &F \times E + Backoff\_delay \end{aligned} \quad (11)$$

The process of deriving *Backoff_delay* is the same, so we do not give details here again.

Performance comparison

$$improvment = \frac{Delay\_original - Delay\_proposed}{Delay\_original} \quad (12)$$

In Table 1, we could see that proposed ertPS has better delay performance. But as the number of station increases, the delay between original scheme and

proposed scheme will become close because our CF slots allocation is based on the ratio of ertPS flows. This improvement is still acceptable.

Theoretical Service flows			Original scheme	Proposed scheme		
				ertPS(QoS)		Others (non-QoS)
Total number of flows	ertPS	Others	Delay (ms)	Delay (ms)	Improvement	Delay (ms)
10	5	5	20.03	13.04	34.90%	14.12
15	5	10	30.24	15.25	49.57%	16.39
20	5	15	37.65	18.48	50.92%	20.15
25	5	20	40.17	27.09	32.56%	32.45
30	5	25	42.99	37.38	13.05%	42.17

Table 1 Theoretical performance comparison.

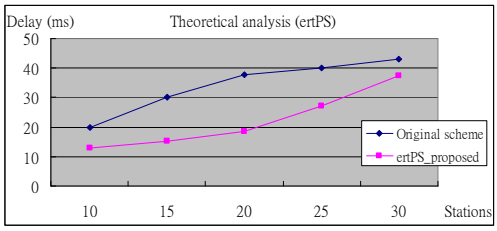


Figure 4 Theoretical analysis of delay comparison.

We analyzed the statistics and illustrate the delay variation in Figure 4, which shows that the proposed scheme has better delay performance for ertPS flows.

4. Simulation and Numerical Results

4.1 Simulation environment

In addition to demonstrating the theoretical performance of our proposed scheme, we use NS-2 (version 2.29) tool [9] with WiMAX PMP module [10]. Our network topology of the simulation is shown in Figure 5 and we focus our simulation on the area which can be controlled by one BS.

Each wireless station either runs bidirectional VoIP traffic with silence suppression or BE traffic which may be an application program such as E-mail or web browser. VoIP traffic supporting silence suppression, with format of G.729 codec, 160 bytes payload and 20ms intervals are used for real-time traffic. G.729 codec could compress 64kbit/s data into 6.4kbit/s to 11.8kbit/s. Other traffics with 512 to 1024 bytes payload are used to simulate the best effort traffic (web browsing or E-mail).

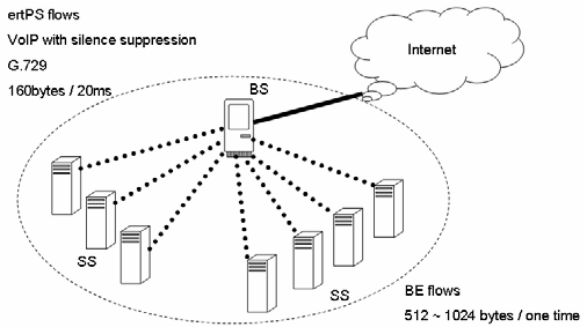


Figure 5 Simulation configuration.

4.2 Numerical results

WiMAX is a new technology in wireless access, so only few modules can be obtained and utilized to simulate the real environments. Our modules just support the WiMAX MAC layer and do not include physical layer, so some simulations do not fit in with actual situation, such as no sufficient bandwidth to deal with large data packets. Even though the PHY module does not satisfy our demands, the simulations still show the results of performance improvement.

Bandwidth request delay

As we described before, bandwidth request represents the requirement of SSs. Bandwidth request delay will cause data packets delay because of insufficient bandwidth for transmission. Of course, it will also lose some throughput.

In our bandwidth request delay simulation, the number of stations is from 10 to 30 and the number of ertPS flows is fixed in 5 to observe delay variation. We are unable to simulate more than 30 stations due to the lack of a suitable WiMAX PHY.

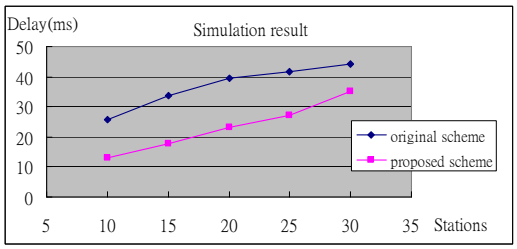


Figure 6 Simulation result of bandwidth request delay-1.

According to the delay data illustrated in Figure 6, our proposed scheme has over 10ms improvement. Delay data is getting close in 30 stations because the allocated CF slots are based on ratio of ertPS flows to the total number of flows (5:25). The fewer the number of CF slots is, the higher the delay will be.

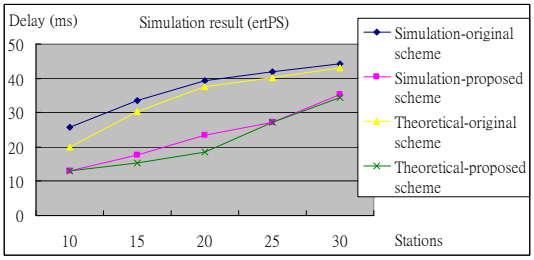


Figure 7 Simulation result of bandwidth request delay-2.

The ertPS delay comparison of four situations which are simulations of the original scheme, the proposed scheme, the theoretical original scheme, and the theoretical proposed scheme, respectively are illustrated in Figure 7. The deviations between two original schemes (theoretical and simulated) are from 2.69% to 21.97% and those between two proposed schemes (theoretical and simulated) are from 0.4% to 22.67%.

In our theoretical analysis, we do not know whether collisions will happen and suppose that our back-off equations are calculated from the first slot. So simulation results will be a little less than theoretical statistics in Figure 7. Simulation value is larger than theoretical value when the number of stations is 20, and we think that the number of stations reaches a bottleneck 20 because it is equal to the number of total contention period slots. If there were collisions, the delay will increase and back-off time is always larger than a frame (back-off window is from 0 to 31 or larger), but our theoretical scheme could not express this situation. When the number of stations is larger than 20, both theoretical and simulation values are close because the back-off window is large enough to handle the situation.

The volume of ertPS flows in the first delay simulation is fixed, so the CF ratio will decrease by increasing the volume of total service flows. Then we try another simulation in which the ratio of ertPS flows is kept at one-third. We could see that the ratio of ertPS flows is close to one-third and the allocated slots are almost 6 or 7. The number of stations is same as before and is no larger than 30.

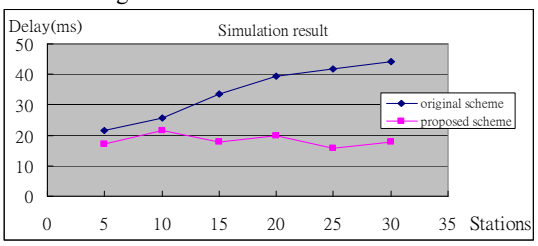


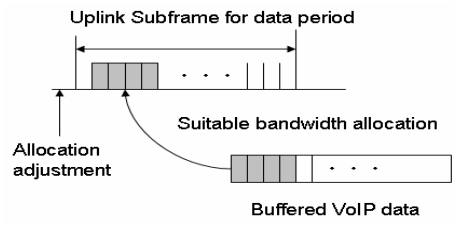
Figure 8 Simulation result of bandwidth request delay-3.

According to Figure 8, the bandwidth request delay of our proposed scheme is pretty close to 20ms. This is because we keep the ratio of ertPS flows to a fixed value and the allocated CF slots are able to handle the demands of those service flows.

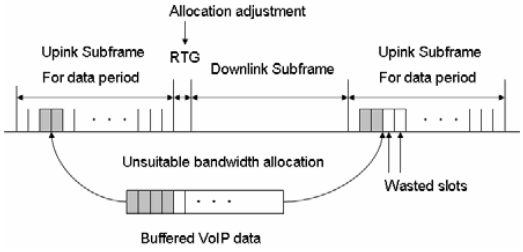
Throughput

VoIP traffic with format of G.729 is used as real-time traffic. Bandwidth request delay relates to throughput, and smaller delay means that the bandwidth allocation is more sensitive. This situation affects the transmission time of VoIP data and we devoted our attention to transmitting VoIP data as soon as possible, so decreasing bandwidth request delay allows VoIP service flows to get proper bandwidth faster.

The difference of data bandwidth allocation between original scheme and the proposed scheme is illustrated in Figure 9. In Figure 9 (a), we do adjustment if buffered VoIP data has suitable bandwidth allocation, data could be transmitted immediately so as not to waste slots. But in Figure 9 (b), heading frame does not have enough bandwidth until the allocation adjustment has been dealt with, so the next frame would be allocated more bandwidth, however, additional slots would be wasted.



(a) Proper bandwidth allocation.



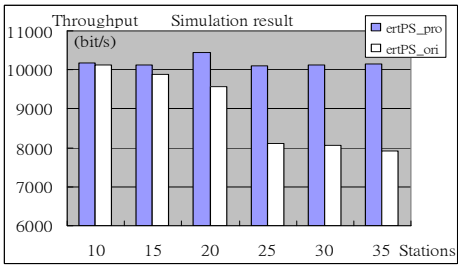
(b) Improper bandwidth allocation.

Figure 9 Comparison of two different bandwidth allocations.

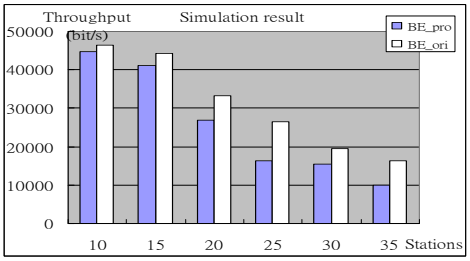
STAs	Average throughput (bit/s)			
	Proposed scheme		Original scheme	
	ertPS	BE	ertPS	BE
10	10133	46424	10169	44764
15	9872	44208	10138	41092
20	9563	33160	10439	26833
25	8104	26410	10102	16330
30	8072	19477	10133	15464
35	7907	16413	10147	9919

Table 2 Simulation result of throughput.

Table 2 illustrates the throughput with different number of stations while our ertPS stations are fixed in 5, so the number of other stations increases. To observe the values, throughput of two different ertPS scheme, which are close to 10k (bit/s) through compression using the G.729 codec.



(a) ErtPS throughput.



(b) BE throughput.

Figure 10 ErtPS and BE throughputs in two schemes.

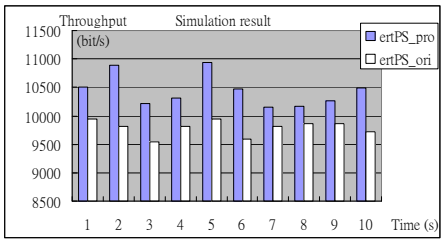


Figure 11 ErtPS throughputs in two different schemes.

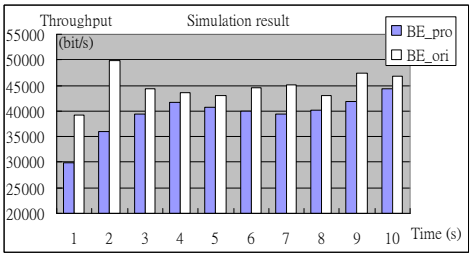


Figure 12 BE throughputs in two different schemes.

BE throughput of our proposed scheme is less than the original one because we neglect their priorities of bandwidth requests. It is still acceptable by its non-QoS characteristic.

Throughputs of two ertPS schemes in Figure 10(a) are averaged values over one second, and the other is the comparison of two BE schemes. Now we observe the throughput during first 10 seconds when the number of stations is 25 and number of ertPS stations is 5. As illustrated in Figure 11 and 12, ertPS throughputs in proposed scheme are higher compared with the original one, while BE throughputs in our proposed scheme are always less than that in the original one.

5. Conclusion and Future Works

In this paper, we analyze the scheme of bandwidth allocation in IEEE 802.16e. BS would allocate suitable bandwidth to SS that has sent the request message; otherwise, data must be buffered waiting for bandwidth. Therefore bandwidth request plays an important role for bandwidth allocation in IEEE 802.16e.

When bandwidth is not sufficient to allocate unicast polling slots for each service flow, multicast polling and broadcast polling should be used. The two polling schemes utilize contention period to let a group of stations contend transmission opportunity of their bandwidth request. Three type service flows, ertPS, nrtPS, and BE must utilize contention period to get transmission slots. In ertPS services, when VoIP flows participate in contention period with other service flows, the VoIP packets delay may be unsettled by the failure of bandwidth request.

We proposed a scheme which let ertPS to have higher priority to transmit its request messages. Our method is to utilize the characteristic in which nrtPS and BE is QoS-free, and we divide original contention period into two different parts: the contention-free (CF) period and contention period. CF period can only be occupied by ertPS requests. Bandwidth requests of ertPS flows will have twice opportunities to be transmitted. A service flow without any allocated CF slot can participate the contention period. But in our scheme, we do not design a precise algorithm to calculate the suitable CF period, but set a threshold which is half of the contention period. We think that a precise algorithm can be designed to utilize slots and avoid wasting of other slots. So we will improve this part in future work.

Based on the simulation results, our scheme reduces the bandwidth request delay and increases certain amount of throughput. Our scheme still could not simulate large number of stations and the environments with wider range, because 802.16 is a new protocol and we could not find a suitable PHY layer module.

We will further improve the scheme by investigating the scenario about bandwidth request. The observation will focus on the QoS degradation induced by collision in contention period. Then we will correct our method to avoid sacrificing the BE service flow transmission opportunities and achieve more improvement in throughput. Also we try to investigate the characteristic of WiMAX PHY layer and integrate it to our scheme.

References

[1] IEEE Std 802.16-2004 (Revision of IEEE Std 802.16-2001).

[2] IEEE Std 802.16e-2005 and IEEE Std 802.16-2004. Cor 1-2005 (Amendment and Corrigendum to IEEE Std 802.16-2004).

[3] "Mobile WiMAX – Part I: A Technical Overview and Performance Evaluation," in WiMAX Forum Feb. 2006.

[4] A. Sayenko, et al., "Ensuring the QoS Requirements in 802.16 Scheduling", ACM MSWiM '06, Oct. 2006, pp. 108-117.

[5] D. Niyato, and E. Hossain, "Queue-aware uplink bandwidth allocation and rate control for polling service in IEEE 802.16 broadband wireless networks", IEEE Transactions on Mobile Computing, June 2006, pp. 668-679.

[6] H. Lee, T. Kwon, and D. Cho, "An enhanced uplink scheduling algorithm based on voice activity for VoIP services in IEEE 802.16d/e system", IEEE Communications Letters, August 2005, pp. 691-693.

[7] R. Iyengar, K. Kar, and B. Sikdar, "Scheduling Algorithms for Point-to-Multipoint Operation in IEEE 802.16 Networks", 4th International Symposium on, April 2006, pp. 1-7.

[8] H. Lee, T. Kwon, and D. Cho, "Performance Analysis of Scheduling Algorithms for VoIP Services in IEEE 802.16e Systems", Vehicular Technology Conference, 2006, pp. 1231-1235.

[9] http://www.isi.edu/nsnam/ns/

[10] http://ndsl.csie.cgu.edu.tw/wimax_ns2.php

[11] C. Y. Liu, "An adaptive bandwidth request scheme for IEEE 802.16e polling services," Master Thesis, National Chiao Tung University, July 2007.