

Bandwidth Recycling in IEEE 802.16 Networks

David Chuck and J. Morris Chang

Abstract—IEEE 802.16 standard was designed to support the bandwidth demanding applications with quality of service (QoS). Bandwidth is reserved for each application to ensure the QoS. For variable bit rate (VBR) applications, however, it is difficult for the subscriber stations (SSs) to predict the amount of incoming data. To ensure the QoS guaranteed services, the SS may reserve bandwidth more than the amount of its transmitting data. As a result, the reserved bandwidth may not be fully utilized all the time. In this paper, we propose a scheme, named *Bandwidth Recycling*, to recycle the unused bandwidth without changing the existing bandwidth reservation. The idea of our scheme is to allow other SSs to utilize the unused bandwidth when it is available. Thus, not only the same QoS guaranteed services can be provided but also the system throughput can be improved. Mathematical analysis and simulation are used to evaluate the proposed scheme. Simulation and analysis results confirm that our proposed scheme can recycle 35 percent of unused bandwidth on average. By analyzing factors affecting the recycling performance, three scheduling algorithms are proposed to improve the overall throughput. The simulation results show that our proposed algorithm can further improve the overall throughput by 40 percent when the network is in the steady state.

Index Terms—WiMAX, IEEE 802.16, bandwidth recycling.

1 INTRODUCTION

THE IEEE 802.16 standards (e.g., 802.16-2004 [1], 802.16e [2]) have received great attention recently. The Worldwide Interoperability for Microwave Access (WiMAX), based on this family of standards, is designed to facilitate services with high transmission rates for data and multimedia applications in metropolitan areas. The physical (PHY) and medium access control (MAC) layers of WiMAX have been specified in the IEEE 802.16 standard. Many advanced communication technologies such as Orthogonal Frequency Division Multiple Access (OFDMA) and multiple-input and multiple-output (MIMO) are embraced in the standards. Supported by these modern technologies, WiMAX is able to provide a large service coverage, high data rates, and QoS guaranteed services. Because of these features, WiMAX is considered to be a promising alternative for last mile broadband wireless access (BWA).

In order to provide QoS guaranteed services, the subscriber station (SS) is required to reserve the necessary bandwidth from the base station (BS) before any data transmissions. In order to serve variable bit rate (VBR) applications, which generate data in variant rates and cannot be modeled accurately, the SS tends to keep the reserved bandwidth to ensure that the QoS guaranteed services can be provided. Thus, it is likely that the amount of data to be transmitted is less than the amount of reserved bandwidth. The reserved bandwidth may not be fully utilized all the time. Although the amount of reserved bandwidth can be adjusted via making bandwidth requests (BRs), the adjusted amount of bandwidth can be applied as early as to the next coming frame. The unused bandwidth in the current frame has no chance to be utilized. Moreover, it is very challenging

to adjust the amount of reserved bandwidth precisely. The SS may be exposed to the risk of degrading the QoS requirement of applications due to the insufficient amount of reserved bandwidth.

To improve the bandwidth utilization while maintaining the same QoS guaranteed services, our research objective is twofold: 1) we do not change the existing bandwidth reservation to maintain the same QoS guaranteed services. 2) Our research work focuses on increasing the bandwidth utilization by utilizing the unused bandwidth. We propose a scheme, named *Bandwidth Recycling*, which recycles the unused bandwidth of each SS while keeping the same QoS guaranteed services and introducing no extra delay. The general concept behind our scheme is straightforward—to allow other SSs to utilize the unused bandwidth left by the current transmitting SS. Since the unused bandwidth is not supposed to occur regularly, our scheme allows SSs with non-real-time applications, which have more flexibility of delay requirements, to recycle the unused bandwidth. Consequently, the unused bandwidth in the *current* frame can be utilized, which is different to the bandwidth adjustment that the amount of bandwidth adjusted can only be enforced as early as in the next coming frame. Moreover, the unused bandwidth is likely to be released temporarily (i.e., only in the current frame) and the existing bandwidth reservation does not change. Therefore, our scheme can improve the overall throughput and bandwidth utilization while providing the same QoS guaranteed services.

According to the IEEE 802.16 standard, SSs scheduled on the uplink (UL) map should have transmission opportunities in the current frame. These SSs are called transmission SSs (TSs) in this paper. The main idea of the proposed scheme is to allow the BS to schedule a backup SS for each TS. The backup SS is assigned to standby for any opportunities to recycle the unused bandwidth of its corresponding TS. We call the backup SS as complementary station (CS). In the IEEE 802.16 standard, BRs are made in per-connection basis. However, the BS allocates bandwidth in per-SS basis. It gives the SS flexibility to allocate the reserved bandwidth to each connection locally.

• The authors are with Iowa State University, Durham Center, Ames, IA 50011. E-mail: {chuang, morris}@iastate.edu.

Manuscript received 1 Oct. 2009; revised 28 Dec. 2009; accepted 28 Feb. 2010; published online 25 June 2010.

For information on obtaining reprints of this article, please send e-mail to: tmc@computer.org, and reference IEEECS Log Number TMC-2009-10-0400. Digital Object Identifier no. 10.1109/TMC.2010.115.

Therefore, the unused bandwidth is defined as the reserved bandwidth which is still available after all connections running on the SS have been served. In our scheme, when a TS has unused bandwidth, it should transmit a special message, called releasing message (RM), to inform its corresponding CS to recycle the unused bandwidth. However, because of the variety of geographical distance between TS and CS and the transmission power of the TS, the CS may not be able to receive the RM sent from the TS. In this case, the benefit of our scheme may be reduced. In this research, we investigate the probability that the CS receives an RM successfully. Our theoretical analysis shows that the CS has at least 42 percent of probability to receive an RM, which is confirmed by our simulation. By further investigating the factors which affect the effectiveness of our scheme, two factors are concluded: 1) the CS cannot receive the RM and 2) the CS does not have non-real-time data to transmit while receiving an RM. To mitigate those factors, additional scheduling algorithms are proposed. Our simulation results show that the proposed algorithms can further improve the average throughput by 40 percent when the network is in the steady state (i.e., 15-75 seconds in our simulation).

The rest of this paper is organized as follows: In Section 2, we provide background information of IEEE 802.16. Motivation and related works are presented in Section 3. Our proposed scheme is presented in Section 4. The analysis of the proposed scheme and simulation results are placed in Section 5 and Section 6. In Section 7, three additional scheduling algorithms are proposed to enhance the performance of the proposed scheme. The simulation results of each scheduling algorithm are shown in Section 8. At the end, the conclusion is given in Section 9.

2 BACKGROUND INFORMATION

The IEEE 802.16 standard specifies three types of transmission mediums supported as the physical layer (PHY): Single Channel (SC), Orthogonal Frequency Division Multiplexing (OFDM), and OFDMA. We assume OFDMA as the PHY in our analytical model since it is employed to support mobility in IEEE 802.16e standard and the scheme working in OFDMA should also work in others. There are four types of modulations supported by OFDMA: BPSK, QPSK, 16-QAM, and 64-QAM.

There are two types of operational modes defined in the IEEE 802.16 standard: point-to-multipoint (PMP) mode and mesh mode. This paper is focused on the PMP mode. In PMP mode, the SS is not allowed to communicate with any other SSs but the BS directly. Based on the transmission direction, the transmissions between BS and SSs can be classified into downlink (DL) and UL transmissions. The former are the transmissions from the BS to SSs. Conversely, the latter are the transmissions in the opposite direction.

There are two transmission modes: Time Division Duplex (TDD) and Frequency Division Duplex (FDD) supported in IEEE 802.16. Both UL and DL transmissions cannot be operated simultaneously in TDD mode but in FDD mode. In this paper, our scheme is focused on the TDD mode. In WiMAX, the BS is responsible for scheduling both UL and DL transmissions. All scheduling behavior is expressed in an MAC frame.

The structure of an MAC frame defined in IEEE 802.16 standard contains two parts: UL subframe and DL subframe. The UL subframe is for UL transmissions. Similarly, the DL subframe is for DL transmissions. In IEEE 802.16 networks, the SS should be coordinated by the BS. All coordinating information including burst profiles and offsets is in the DL and UL maps, which are broadcasted at the beginning of an MAC frame.

The IEEE 802.16 network is connection-oriented. It gives the advantage of having better control over network resource to provide QoS guaranteed services. In order to support wide variety of applications, the IEEE 802.16 standard classifies traffics into five scheduling classes based on different QoS requirements: 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). When serving applications, the SS classifies each application into one of the scheduling classes and establishes a connection with the BS based on its scheduling class. The BS assigns a connection ID (CID) to each connection. When a connection needs more bandwidth, the SS requests bandwidth based on its CID via sending a BR. When receiving a BR, the BS can either grant or reject the request depending on its available resources and scheduling policies.

There are two types of BRs defined in the IEEE 802.16 standard: incremental and aggregate BRs. Incremental BRs allow the SS to indicate the amount of extra bandwidth required for a connection. Thus, the amount of reserved bandwidth can only be increased via incremental BRs. On the other hand, the SS specifies the current state of queue for the particular connection via a aggregate request. The BS resets its perception of that service's needs upon receiving the request. Consequently, the reserved bandwidth may be decreased.

3 MOTIVATION AND RELATED WORK

Bandwidth reservation allows IEEE 802.16 networks to provide the QoS guaranteed services. The SS reserves the required bandwidth before any data transmissions. Due to the nature of VBR applications, it is very difficult for the SS to request the bandwidth accurately to ensure the QoS requirement of applications. It is possible that the amount of reserved bandwidth is more than the number of data that the SS transmits. Therefore, the reserved bandwidth cannot be fully utilized. Although making BRs is the scheme defined in the standard to help the SS adjust the amount of reserved bandwidth; however, the updated amount of reserved bandwidth is applied as early as to the next coming frame. The unused bandwidth in the current frame still cannot be utilized. In our scheme, the SS is able to release its unused bandwidth temporally (i.e., only in the current frame). Another SS which is preassigned by the BS tries to utilize this unused bandwidth. This can improve the bandwidth utilization, which leads to better system throughput. Moreover, since the existing bandwidth reservation is not changed, the same QoS guaranteeing service can be provided and no extra delay is introduced.

Many research works dealing with the improvement of bandwidth utilization and system throughput have been proposed in the literature. In [4], a dynamic resource reservation mechanism is proposed. It can dynamically

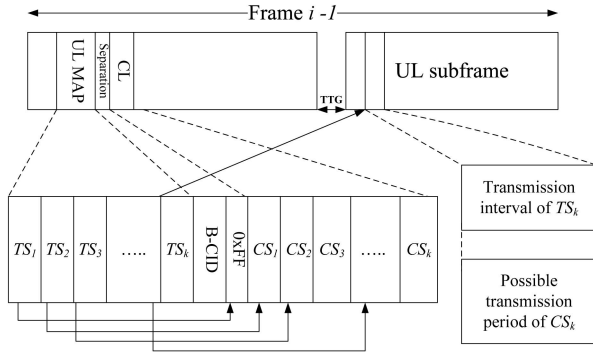


Fig. 1. The mapping relation between CSs and TSs in an MAC frame.

change the amount of reserved resource depending on the actual number of active connections. The investigation of dynamic bandwidth reservation for hybrid networks is presented in [3]. The authors evaluate the performance and effectiveness for the hybrid network and find efficient methods to ensure optimum reservation and utilization of bandwidth while minimizing signal blocking probability and signaling cost. In [5], the authors enhanced the system throughput by using concurrent transmission in mesh mode. The authors in [6] proposed a new QoS control scheme by considering MAC-PHY cross-layer resource allocation. A dynamic bandwidth request allocation algorithm for real-time services is proposed in [7]. The authors predict the amount of bandwidth to be requested based on the information of the backlogged amount of traffic in the queue and the rate mismatch between packet arrival and service rate to improve the bandwidth utilization. The research works listed above improve the performance by predicting the traffic coming in the future. Instead of prediction, our scheme can allow SSs to accurately identify the portion of unused bandwidth and provides a method to recycle the unused bandwidth. It can improve the utilization of bandwidth while keeping the same QoS guaranteed services and introducing no extra delay.

4 PROPOSED SCHEME

The objectives of our research are twofold: 1) The same QoS guaranteed services are provided by maintaining the existing bandwidth reservation and 2) the bandwidth utilization is improved by recycling the unused bandwidth. To achieve these objectives, our scheme named as *Bandwidth Recycling* is proposed. The main idea of the proposed scheme is to allow the BS to preassign a CS for each TS at the beginning of the current frame. The CS waits the possible opportunities to recycle the unused bandwidth of its corresponding TS in this frame. The CS information scheduled by the BS is resided in a list, called complementary list (CL). The CL includes the mapping relation between each pair of preassigned CS and TS. As shown in Fig. 1, each CS is mapped to at least one TS. The CL is broadcasted followed by the UL map. For the backward compatibility, a broadcast CID (B-CID) is attached in front of the CL. Moreover, a stuff byte value (SBV) is transmitted followed by the B-CID to distinguish the CL from other broadcast DL transmission intervals.

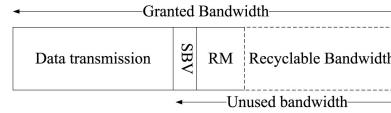


Fig. 2. Messages to release the unused bandwidth within a UL transmission interval.

HT = 0	EC	TYPE	Rev	CI	EKS	Rev	LEN MSB
		LEN LSB					CID MSB
		CID LSB					HCS

HT : Header Type CI : CRC Indicator MSB: Most Significant Bit
 EC : Encryption Control EKS: Encryption Key Sequence LSB: Least Significant Bit
 CID : Connection ID LEN : Length HCS: Header Check Sequence

Fig. 3. The format of RM.

The UL map including burst profiles and offsets of each TS is received by all SSs within the network. Thus, if an SS is scheduled on both UL map and CL, the necessary information (e.g., burst profile) residing in the CL may be reduced to the mapping information between the CS and its corresponding TS. The BS only specifies the burst profiles for the SSs which are only scheduled on the CL. For example, as shown in Fig. 1, CS_j is scheduled as the corresponding CS of TS_j , where $1 \leq j \leq k$. When TS_j has unused bandwidth, it performs our protocol introduced in Section 4.1. If CS_j receives the message sent from TS_j , it starts to transmit data by using the burst profile decided by the BS. The burst profile of a CS can be resided on either the UL map if the CS is also scheduled on the UL map or the CL if the CS is only scheduled on CL.

Our proposed scheme is presented into two parts: the protocol and scheduling algorithm. In the protocol, we introduce how the TS identifies the unused bandwidth and gives recycling opportunities to its corresponding CS. The scheduling algorithm helps the BS to schedule a CS for each TS.

4.1 Protocol

According to the IEEE 802.16 standard, the allocated space within a data burst that is unused should be initialized to a known state. Each unused byte should be set as a padding value (i.e., 0xFF), called SBV. If the size of the unused region is at least the size of an MAC header, the entire unused region is suggested to be initialized as an MAC PDU. The padding CID (value of 0xFFFE) is used in the CID field of the MAC PDU header. In this research, we intend to recycle the unused space for data transmissions.

Instead of padding all portion of the unused bandwidth in our scheme, a TS with unused bandwidth transmits only an SBV and an RM shown in Fig. 2. The SBV is used to inform the BS that there are no more data coming from the TS. On the other hand, the RM is composed of a generic

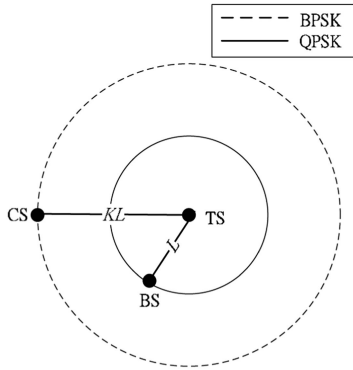


Fig. 4. An example of corresponding locations of TS, BS, and CS.

MAC PDU with no payload (6 bytes), as shown in Fig. 3. The mapping information between CL and UL map is based on the basic CID of each SS. The CID field in RM should be filled by the basic CID of the TS.

Since there is an agreement of modulation for transmissions between TS and BS, the SBV can be transmitted via this agreed modulation. However, there are no agreed modulations between TS and CS. Moreover, the transmission coverage of the RM should be as large as possible in order to maximize the probability that the RM is able to be received successfully by the CS. To maximize the transmission coverage of the RM, one possible solution is to increase the transmission power of the TS while transmitting the RM. However, power may be a critical resource for the TS and should not be increased dramatically. Therefore, under the condition of without increasing the transmission power of the TS, the RM should be transmitted via BPSK which provides the largest coverage among all modulations supported in the IEEE 802.16 standard.

For example, Fig. 4 illustrates the physical location of the BS, TS, and CS, respectively. The solid circle represents the coverage of QPSK which is the modulation for the data transmissions between BS and TS. When the TS has unused bandwidth, it transmits the SBV via this modulation (i.e., QPSK) to inform the BS that there are no more data coming from the TS. From the figure, it is easy to observe that the corresponding CS is out of QPSK coverage. In order to maximize the coverage of the RM under the condition of without increasing the transmission power of the TS, the TS transmits the RM via BPSK which coverage is represented by the dashed circle. The radius of the dashed circle is KL , where L is the distance between TS and BS and K is the ratio of transmission range of BPSK to the transmission range of QPSK depending on the transmission power. Assume that all channels are in good condition. As long as the CS is within the coverage of BPSK, it can receive the RM successfully and start to recycle the unused bandwidth of the TS.

Since both UL map and CL can be received by the CS, the CS knows the UL transmission period of its corresponding TS. This period is called the UL transmission interval. The CS monitors this interval to see if an RM is received from its corresponding TS. Once an RM is received, the CS starts to recycle the unused bandwidth by using the burst profile residing in either UL map (if the CS is scheduled on the UL map as well) or CL (if the CS is only scheduled on the CL), until using up the rest of the TS's transmission interval. If

the CS does not have any data to transmit, it simply pads the rest of the transmission interval.

4.2 Scheduling Algorithm

Assume that Q represents the set of SSs which serve non-real-time connections (i.e., nrtPS or BE connections) and T is the set of TSs. Due to the feature of TDD that the UL and DL operations cannot be performed simultaneously, we cannot schedule the SS which UL transmission interval is overlapped with the target TS.

For any TS, S_t , let O_t be the set of SSs which UL transmission interval overlaps with that of S_t in Q . Thus, the possible corresponding CS of S_t must be in $Q - O_t$. All SSs in $Q - O_t$ are considered as candidates of the CS for S_t . A scheduling algorithm, called *Priority-based Scheduling Algorithm (PSA)*, shown in Algorithm 1 is used to schedule an SS with the highest priority as the CS. The priority of each candidate is decided based on the scheduling factor (SF) which is the ratio of the current requested bandwidth (CR) to the current granted bandwidth (CG). The SS with higher SF has more demand on the bandwidth. Thus, we give the higher priority to those SSs. The highest priority is given to the SSs with zero CG. Non-real-time connections include nrtPS and BE connections. The nrtPS connections should have higher priority than the BE connections because of the QoS requirements. The priority of candidates of CSs is concluded from high to low as: nrtPS with zero CG, BE with zero CG, nrtPS with nonzero CG, and BE with nonzero CG. If there are more than one SS with the highest priority, we pick one with the largest CR as the CS in order to decrease the probability of overflow.

Algorithm 1. Priority-based Scheduling Algorithm

Input: T is the set of TSs scheduled on the UL map.

Q is the set of SSs running non-real-time applications.

Output: Schedule CSs for all TSs in T .

For $i = 1$ to $\|T\|$ **do**

a. $S_t \leftarrow TS_i$.

b. $Q_t \leftarrow Q - O_t$.

c. Calculate the SF for each SS in Q_t .

d. **If** Any SS $\in Q_t$ has zero granted bandwidth,

If Any SSs have nrtPS traffics and zero granted bandwidth,

Choose one running nrtPS traffics with the largest CR.

else

Choose one with the largest CR.

else

Choose one with largest SF and CR.

e. Schedule the SS as the corresponding CS of S_t .

End For

5 ANALYSIS

The percentage of potentially unused bandwidth occupied in the reserved bandwidth is critical for the potential performance gain of our scheme. We investigate this percentage on VBR traffics which is one of popular traffic type used today. Additionally, in our scheme, each TS should transmit an RM

to inform its corresponding CS when it has unused bandwidth. However, the transmission range of the TS may not be able to cover the corresponding CS. It depends on the location and the transmission power of the TS. It is possible that the unused bandwidth cannot be recycled because the CS may not be able to receive the RM. Therefore, the benefit of our scheme may be reduced. In this section, we analyze mathematically the probability of a CS to receive an RM successfully. Obviously, this probability affects the bandwidth recycling rate (BBR). BBR stands for the percentage of the recycled unused bandwidth. Moreover, the performance analysis is presented in terms of throughput gain (TG). At the end, we evaluate the performance of our scheme under different traffic load. All analytical results are validated by the simulation in Section 6.

5.1 Analysis of Potential Unused Bandwidth

Based on the traffic generation rate, the applications can be classified into two types: constant bit rate (CBR) and VBR. Since CBR applications generate data in a constant rate, SSs rarely adjust the reserved bandwidth. As long as the reasonable amount of bandwidth is reserved, it is hard to have unused bandwidth in this type of applications. Therefore, our scheme has very limited benefit on CBR traffic. However, VBR applications generate data in a variable rate. It is hard for an SS to predict the amount of incoming data precisely for requesting the appropriate bandwidth to satisfy the QoS requirements. Thus, in order to provide QoS guaranteed services, the SS tends to keep the amount of reserved bandwidth to serve the possible bursty data arrived in the future. The reserved bandwidth may not be fully utilized all the time. Our analysis focuses on investigating the percentage of potentially unused bandwidth of VBR traffics.

In our traffic model based on [8], the time interval between arriving packets of the VBR traffic is considered as exponential distribution. The steady-state probability of the traffic model can be characterized by Poisson distribution. Let λ and λ_{max} be the mean and maximal amount of data arriving in a frame, respectively. Suppose X represents the amount of data arriving in a frame and $p(X)$ is the probability of X amount of data arriving in a frame, where $0 \leq X \leq \lambda_{max}$.

When the SS intends to establish a new connection with the BS, this connection must pass the admission control in order to make sure that the BS has enough resource to provide QoS guaranteed services. The policy can be considered as a set of predefined QoS parameters such as minimum reserved traffic rate (R_{min}), maximum sustained rate (R_{max}), and maximum burst size (W_{max}) [9], [10]. In our analytic model, the BS initially assigns the bandwidth, B , to each connection. The BS guarantees to support the bandwidth until reaching R_{min} and optionally to reach R_{max} . Suppose D_f represents the frame duration and W is the assigned bandwidth per frame (in terms of bytes). Because of the admission control policy, the burst size that the BS schedules in each frame cannot be larger than W_{max} . The relation between W and B can be formulated as

$$W = BD_f \leq W_{max}. \quad (1)$$

Suppose X_{i-1} represents the amount of data arriving in the frame $i-1$ (in terms of bytes), where $1 \leq i \leq N-1$ and N is the total number of frames we analyze. If we have unused bandwidth in frame i , then the amount of data in queue must be less than the number of assigned bandwidth. By considering the interframe dependence (i.e., the number of data changed in the previous frame affects the number of data in queue in the current frame), it can be represented as the following condition:

$$X_{i-1} < W_i - \max\{0, Q_{i-1} - W_{i-1}\}, \quad (2)$$

where Q_{i-1} is the amount of data stored in queue before transmitting frame $i-1$. W_i and W_{i-1} are the amount of bandwidth assigned in frame i and $i-1$, respectively. Again, both W_i and W_{i-1} are at most W_{max} . $\max\{0, Q_{i-1} - W_{i-1}\}$ represents the amount of queued data arriving before frame $i-1$.

As mentioned, X_{i-1} is the amount of data arriving in the frame $i-1$. Thus, X_{i-1} must be nonnegative. Consequently, the probability of having unused bandwidth in frame i , $P_u(i)$, is derived as

$$P_u(i) = \int_0^{X_{i-1}} p(X) dX. \quad (3)$$

Thus, the expected amount of unused bandwidth in frame i , $E(i)$, can be derived as

$$E(i) = \int_0^{X_{i-1}} Xp(X) dX. \quad (4)$$

Finally, by summing the expected unused bandwidth in all frames, the ratio of the total potentially unused bandwidth to total reserved bandwidth in N frames, R_u , can be presented as

$$R_u = \frac{\sum_{i=0}^{N-1} E(i)}{\sum_{i=0}^{N-1} W_i}. \quad (5)$$

5.2 The Probability of RMs Received by the Corresponding CSs Successfully

Assume that a BS resides at the center of a geographical area. There are n SSs uniformly distributed in the coverage area of BS. Since PMP mode is considered, the transmissions only exist between BS and SSs. Moreover, each SS may be in different locations. The transmission rate of each SS may be variant depending on the PHY transmission technology and transmission power. For a given SS, S_t , let $R_t^{(B)}$, $R_t^{(Q)}$, $R_t^{(16)}$, and $R_t^{(64)}$ denote as the transmission range of BPSK, QPSK, 16-QAM, and 64-QAM, respectively. In our scheme, the RM should be transmitted via the most robust modulation (i.e., BPSK) since it has the largest coverage of RMs among all modulations supported by the IEEE 802.16 standard when the transmission power is not adjusted. Based on the fixed transmission power, the relation of transmission range between modulations can be expressed as

$$R_t^{(B)} = k_t^{(Q)} R_t^{(Q)} = k_t^{(16)} R_t^{(16)} = k_t^{(64)} R_t^{(64)},$$

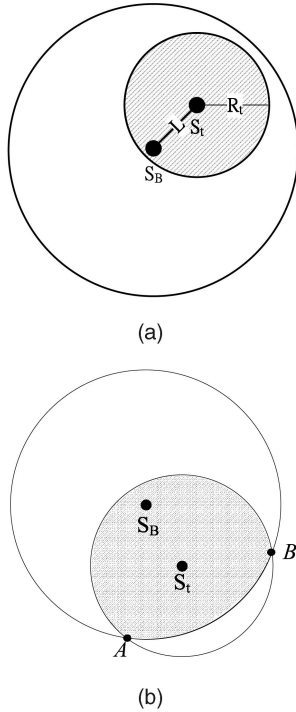


Fig. 5. Possible geographical relationship between S_t and S_B . (a) All coverage of S_t is within the coverage of S_B . (b) The coverage of S_t is partially within the coverage of S_B .

where $k_t^{(Q)}$, $k_t^{(16)}$, and $k_t^{(64)}$ are constants depending on the transmission power of S_t and $k_t^{(64)} \geq k_t^{(16)} \geq k_t^{(Q)} \geq 1$. Again, the RM should be transmitted via BPSK. In the rest of the paper, we use R_t to represent the BPSK transmission range of S_t . Moreover, S_B and R are denoted the BS and its transmission range of BPSK, respectively.

Each TS may use different transmission power to communicate with the BS, depending on the distance between them and the modulation used for communications. In our scheme, we do not intend to increase the transmission power of the TS. Therefore, the RM should be transmitted via BPSK which has the largest coverage among all modulations. However, the transmission coverage of the RM may not be able to cover the whole service area of S_B . Consequently, it is possible that the CS cannot receive the RM. Furthermore, it is worth noticing that the location of the TS also affects the probability of a CS to receive the RM. Therefore, we must analyze the probability that a CS can receive an RM from its corresponding TS successfully.

From the UL map and CL, the CS can obtain the UL transmission interval of its corresponding TS. Thus, the CS starts to expect an RM at the beginning of the UL transmission interval of its corresponding TS. Additionally, since SSs are randomly distributed in the service area of S_B , the probability of a CS to receive an RM is equivalent to the transmission coverage of an RM overlapping with the service coverage of S_B . We analyze the average probability that the CS can receive an RM successfully.

For any TS S_t , suppose S_j is denoted as the CS of S_t . The relationship between S_t and S_B can be classified into two categories based on the location of S_t : 1) all coverage of S_t is within the service coverage of S_B , as shown in Fig. 5a, and 2) only part of the coverage of S_t is within the service

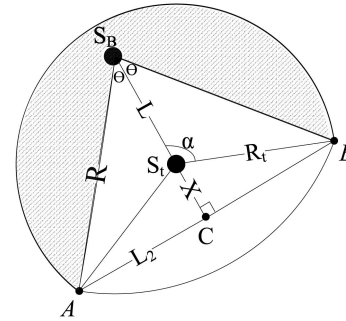


Fig. 6. Both S_B and S_t are in the same side of \overline{AB} .

coverage of S_B , as shown in Fig. 5b. The coverage of S_t means the maximal coverage of RMs transmitted by S_t . The analysis of each category is presented as follows:

5.2.1 The Coverage of S_t is within the Coverage of S_B

In this category, all coverage of S_t is within the service area of S_B . The coverage of S_t , denoted as A_{in} , can be derived as

$$A_{in} = \pi R_t^2. \quad (6)$$

The probability of S_j receiving the RM, denoted as $P_c(t)$, is the same as the ratio of converges of S_t to S_B :

$$P_c(t) = \frac{R_t^2}{R^2}. \quad (7)$$

Moreover, the coverage of the two stations (S_t and S_B) must intersect on no more than one point. Suppose L represents the distance between S_t and S_B . The condition to have this type of situation can be expressed in terms of L :

$$L \leq R - R_t. \quad (8)$$

Because R_t represents the BPSK transmission range of S_t , we can have

$$R_t = KL, \quad (9)$$

where K is a constant depending on the transmission power and modulation that S_t uses to communicate with the S_B . By combining (8) and (9), S_t belongs to this category if:

$$L \leq \frac{R}{K+1}. \quad (10)$$

By calculating the area with radius L , the probability of S_t within this category, $P_{oc}(t)$, is

$$P_{oc}(t) = \frac{1}{(K+1)^2}. \quad (11)$$

5.2.2 The Coverage of S_t Is Partially within the Coverage of S_B

The boundary of S_t intersects with the boundary of S_B at two points A and B , as shown in Fig. 5b. Based on the location of S_t , we can classify into two cases:

1. Both S_t and S_B are on the same side of \overline{AB} .

Fig. 6 illustrates the RM coverage of S_t overlapping with the service area of S_B and both stations reside on the same side of \overline{AB} . Because of the limited space,

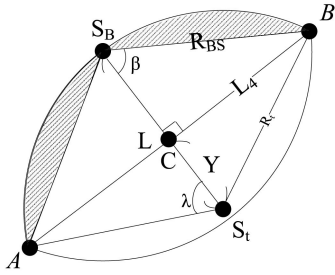


Fig. 7. S_B and S_t are in each side of \overline{AB} .

the calculation is omitted from this paper. The total area A_{total} can be presented as

$$A_{total} = R^2\theta + R_i^2\alpha - LL_2. \quad (12)$$

Consequently, the probability of S_j receiving the RM, $P_s(t)$, can be derived as

$$P_s(t) = \frac{R^2\theta + R_i^2\alpha - LL_2}{\pi R^2}. \quad (13)$$

In this case, the borders of both S_t and S_B coverage must intersect on two points. From (10), L must be longer than $\frac{R}{K+1}$ which is the lower bound of this case. Moreover, since both S_B and S_t must reside on the same side of \overline{AB} , L must be no longer than the shortest distance from BS to \overline{AB} . Thus, we can derive the upper bound of L as

$$L \leq \frac{R}{\sqrt{1+K^2}}. \quad (14)$$

By calculating the ring area between lower bound and upper bound, the probability of S_t in this case, $P_{os}(t)$, can be derived as

$$P_{os}(t) = \frac{2K}{(K+1)^2(1+K^2)}. \quad (15)$$

2. S_B and S_t are on different sides of \overline{AB} .

Fig. 7 illustrates the overlapping coverage of S_t and S_B . Each of them is located on one side of \overline{AB} . The total area A'_{total} that S_j can receive the RM is

$$A'_{total} = R^2\beta + R_i^2\lambda - LL_4. \quad (16)$$

Therefore, the probability of S_j receiving RMs can be derived as

$$P_e(t) = \frac{R^2\beta + R_i^2\lambda - LL_4}{\pi R^2}. \quad (17)$$

Since each of S_t and S_B is in one side of \overline{AB} , the distance between S_t and S_B must be longer than the shortest distance from S_B to \overline{AB} . From (14), we can obtain that L must be longer than $\frac{R}{\sqrt{1+K^2}}$ which is the lower bound of this case. Moreover, S_t needs to stay in the service area of S_B . Thus, L cannot be longer than R . By calculating the ring area between the lower bound and upper bound of L , the probability of S_t belonging to this case, $P_{oc}(t)$, can be derived as

$$P_{oc}(t) = \frac{K^2}{1+K^2}. \quad (18)$$

From the two categories shown above, the probability of S_j to receive an RM from S_t can be concluded as

$$P_t(t) = P_e(t)P_{oc}(t) + P_s(t)P_{os}(t) + P_c(t)P_{oc}(t). \quad (19)$$

Consequently, on average, the probability of a CS to receive the RM from its corresponding TS can be derived as

$$P_t = \frac{\sum_{t=1}^{\|T\|} P_t(t)}{\|T\|}, \quad (20)$$

where T is the set of all TSs.

5.3 Performance Analysis of the Proposed Scheme

Assume that Q_n represents a set of SSs running non-real-time connections and Q_{CL} is a set of SSs in Q_n scheduled as CSs. Thus, $\|Q_{CL}\|$ is at most $\|T\|$, where T is the set of all TSs. For any SS, $S_n \in Q_n$, the probability of S_n scheduled on the CL, $P_{CL}(n)$, can be derived as

$$P_{CL}(n) = \begin{cases} \frac{\|Q_{CL}\|}{\|Q_n\|}, & \|Q_n\| \geq \|Q_{CL}\|, \\ 1, & \text{otherwise.} \end{cases} \quad (21)$$

It is possible that the CS fails to recycle the unused bandwidth due to the lack of no-real-time data to be transmitted. Thus, it is necessary to analyze this probability. Suppose Y_{i-1} is the amount of non-real-time data arriving in frame $i-1$. The amount of bandwidth assigned in frame i and $i-1$ is denoted as W_i^{nrt} and W_{i-1}^{nrt} , respectively. Obviously, both W_i^{nrt} and W_{i-1}^{nrt} cannot be larger than W_{max}^{nrt} , where W_{max}^{nrt} is the maximum burst size. If the CS can recycle the unused bandwidth in frame i , then the amount of data in queue must be more than W_i^{nrt} . In the consideration of interframe dependence, it can be expressed as the following condition:

$$Y_{i-1} > W_i^{nrt} - \max\{0, Q_{i-1}^{nrt} - W_{i-1}^{nrt}\}, \quad (22)$$

where $\max\{0, Q_{i-1}^{nrt} - W_{i-1}^{nrt}\}$ is the amount of queued data arriving before frame $i-1$.

Since Y_{i-1} cannot be negative, the probability of the CS, denoted as S_u , which has data to recycle the unused bandwidth can be obtained as

$$P_u(u) = \int_{Y_{i-1}}^{\lambda_{max}^{nrt}} P(X)dX, \quad (23)$$

where λ_{max}^{nrt} is the maximal amount of non-real-time data arriving in a frame.

A CS which recycles the unused bandwidth successfully while receiving an RM must be scheduled on the CS and have non-real-time data to be transmitted. From (21) and (23), the probability that a CS satisfies these two conditions can be derived as

$$P_r = \frac{\sum_{j=1}^{\|Q_n\|} P_u(j)(P_{CL}(j))}{\|Q_n\|}. \quad (24)$$

If the CS recycles the unused bandwidth successfully, then it must meet the three conditions: 1) an RM must be received, 2) this SS must be scheduled on the CL, and 3) the CS must have data to recycle the unused bandwidth. From (20) and (24), the recycling rate, defined as the average probability that a CS recycles the unused bandwidth successfully, can be obtained as

$$P_{recycle} = P_r P_t. \quad (25)$$

Suppose B_g is the total bandwidth in the system and the unused bandwidth of the system is B_w . By (25), The total throughput gain TG can be derived as

$$TG = \frac{P_{recycle} B_w}{B_g - B_w}. \quad (26)$$

Delay is a critical factor affecting the QoS of services. In our scheme, we preserve the existing bandwidth reservation. Moreover, the CS cannot recycle the bandwidth until receiving the RM which is sent by the TS. Therefore, *Bandwidth Recycling* does not affect any data transmissions operated by the TS and it does not introduce any extra delay.

5.4 Overhead Analysis of the Proposed Scheme

The overhead introduced by our scheme resides in both DL and UL subframes. In DL subframe, the separation and CL are considered as the overhead. As shown in Fig. 1, the separation contains a broadcast CID (B-CID) and an SBV (0xFF). It costs 3 bytes of overhead (16 bits for B-CID and 1 byte for SBV). In addition, the CL is composed of the CL information elements (CL-IEs). The CL-IE contains the basic CID of the CS. If the CS is not scheduled on the UL map, the burst profile and offset must be specified in the CL-IE of this CS. Therefore, the size of CL-IE is at most the size of UL-MAP IE which is 7 bytes defined in the IEEE 802.16 standard. In summary, the total overhead in a DL subframe can be concluded as

$$OH_{DL} \leq 3 + 7B_{TS}, \quad (27)$$

where B_{TS} is the number of TSs scheduled on the UL map.

According to the IEEE 802.16 standard, the SBV is inevitable when the SS has unused bandwidth. Therefore, only RMs are considered as the overhead in UL subframe. The RM is used for a TS to inform its corresponding CS to recycle the unused bandwidth. Therefore, each TS can transmit at most one RM in each UL subframe. An RM is composed of a generic MAC Header (GMH). The size of a GMH is 6 bytes defined in the IEEE 802.16 standard. Thus, the total overhead in a UL subframe is calculated as

$$OH_{UL} \leq 6B_{TS}, \quad (28)$$

where B_{TS} is the number of TSs scheduled on the UL map. From (27) and (28), the total overhead introduced by our scheme in an MAC frame is concluded as

$$OH = OH_{DL} + OH_{UL} \leq 3 + 7B_{TS} + 6B_{TS}. \quad (29)$$

5.5 Performance Analysis of the Proposed Scheme under Different Traffic Load

The traffic load in a network may vary at different time points. Based on this, the network status can be classified into four stages: light, moderate, heavy, and fully loaded. The performance of the proposed scheme may be variant in different stages. We investigate the performance of our scheme in each stage. Suppose B_{all} represents the total bandwidth supported by the BS. Assume that B_{rt} represents the bandwidth reserved by real-time connections and BR_{rt} is the amount of additional bandwidth requested by them via BRs. Similarly, B_{nrt} represents the bandwidth assigned to non-real-time connections and BR_{nrt} is the amount of additional bandwidth requested by them. The investigation of our scheme in each stage is shown as follows: All investigations are validated via simulation in Section 6.

1. **Stage 1 (light load).** This stage is defined as that the total demanding bandwidth of SSs is much less than the supply of the BS. The formal definition can be expressed as

$$B_{all} \gg B_{rt} + B_{nrt} + BR_{rt} + BR_{nrt}.$$

Since all BRs are granted in this stage, the BS schedules the CS randomly. Moreover, every SS receives its desired amount of bandwidth. Therefore, for any given CS, S_u , the probability to have data to recycle the unused bandwidth, derived from (23), is small. It leads to low P_r (from (24)). Therefore, the probability that the CS recycles the unused bandwidth successfully is small and the throughput gain of our scheme is not significant.

2. **Stage 2 (moderate load).** This network stage is defined as equal demand and supply of bandwidth, i.e.,

$$B_{all} = B_{rt} + B_{nrt}.$$

In this stage, BS can satisfy the existing demand but does not have available resource to admit new BRs. Since the currently desired bandwidth of every SS can be satisfied, the probability of CS to recycle the unused bandwidth (23) may be higher than the stage 1 but still limited. Based on (24), (25), and (26), the throughput gain is still insignificant.

3. **Stage 3 (heavy load).** This stage is defined as that the BS can satisfy the demand of real-time connections, but does not have enough bandwidth for the non-real-time connections. However, there are no rejected BRs in this stage. We can express this in terms of formulation as

$$B_{all} = B_{rt} + \kappa B_{nrt},$$

where $0 \leq \kappa < 1$. Since the bandwidth for non-real-time connections has been shrunk, there is a high probability that the CS accumulates non-real-time data in queue. It leads to higher P_r and $P_{recycle}$. Thus, the throughput gain can be more significant than Stages 1 and 2.

4. **Stage 4 (full load).** This stage describes a network with the heaviest traffic load. The difference between stages 3 and 4 is that there are rejected BRs in stage 4. It means that the probability of SSs

TABLE 1
The System Parameters Used in Our Simulation

Parameters	Value
Node number	51 (including BS)
Frame duration	20MS
UL/DL subframe duration	10MS
Modulation scheme	BPSK, QPSK, 16QAM, 64QAM
DCD/UCD broadcast interval	5S
TTG/RTG	10US
SS transition gap (SSTG)	4US

accumulating non-real-time data in queue is much higher than the one in Stage 3. Therefore, both P_r and $P_{recycle}$ are significantly high. Our scheme can achieve the best performance in this stage.

5.6 Trade-Off

In the IEEE 802.16 standard, the SS can adjust the amount of reserved bandwidth via BRs. In this section, we analyze the performance between the proposed scheme and the scheme with BRs. However, there are no rules specified in the standard to tell the SS when to adjust the amount of reserved bandwidth. The objective of this paper is to improve the bandwidth utilization and system throughput. We define a case, named Case with BRs, that each SS requests bandwidth for each connection in every frame based on the queued data. The unicast polling opportunity is given to each connection in every frame for making BRs.

In this case, in each frame, the SS always asks the amount of bandwidth as the number of data it will transmit. Therefore, the amount of unused bandwidth in this case is very limited. However, the SS has to transmit a BR for every connection in every frame. Moreover, according to the IEEE 802.16 standard, the BR is made in per connection basis. Suppose there are m connections running on an SS. The SS has to send m BRs which are $19m$ bytes (considering stand-alone bandwidth requests) in each frame. The overhead is dramatically large in this case. Since the size of UL subframe is limited in each frame, the throughput for transmitting real data (i.e., eliminating the overhead) may not be high. On the other hand, in the proposed scheme, the overhead that each SS transmits is a constant (6 bytes for an RM) which is much smaller than $19m$ bytes.

Since the CS needs to stay in active in order to listen to a possible RM from the corresponding TS, the CS cannot enter into sleep mode for power conservation. On the other hand, the probability of a CS to recycle the unused bandwidth decreases if a sleeping SS is scheduled as the CS. Therefore, there is a trade-off between the benefit of the proposed scheme and power conservation. If the CS does not enter into sleep mode, obviously, it can always listen to a possible RM sent from the corresponding TS. On the other hand, it enters into sleep mode. The SS switches its state between active and inactive. As described in the IEEE 802.16e standard, the BS has the information of available and unavailable period of the SS. Thus, the BS should avoid to schedule an SS which is in unavailable period as a CS. Furthermore, if the BS schedules an inactive SS as a CS, the whole network still operates successfully but the benefit of the proposed scheme is reduced.

TABLE 2
The Traffic Model Used in the Simulation

Application	VoIP	Multimedia	HTTP	FTP
Traffic type	CBR	VBR	VBR	VBR
Scheduling class	UGS	rtPS	BE	nrtPS
Start Time(sec.)	m*	m*	m*	m*
End Time(sec.)	n*	n*	100	100
Mean Packet Size	512	z*	z*	z*
Mean Bit Rate	12.2kbps	2Mbps	2kbps	50Mbps
Max burst Size (Byte)	31	7.5k	10	1500k
Packet Size	Fixed	P*	P*	P*
Packet Arrival Rate	Fixed	E*	E*	E*

Note: m* is a random number between 0 and 15.
n* is a random number between 75 and 100.
z* is a random number between 512 and 1024 bytes
P* stands for Poisson distribution
E* stands for Exponential distribution

6 SIMULATION RESULTS

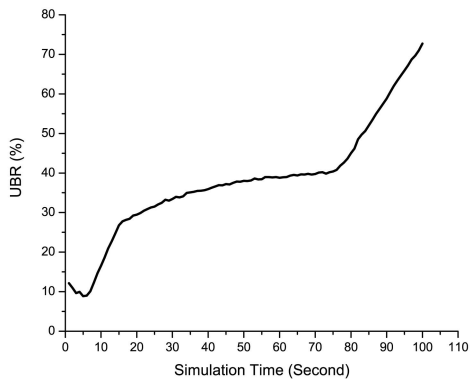
Our simulation is conducted by using Qualnet 4.5 [11], a commercially available network simulator. In this section, we first present our simulation model followed by introducing the definition of performance metrics used for measuring the network performance. The simulation results are shown as the third part of this section. At the end, we provide the validation of theoretical analysis and simulation results.

6.1 Simulation Model

Our simulation model is composed of one BS residing at the center of geographical area and 50 SSs uniformly distributed in the service coverage of BS. The parameters of PHY and MAC layers used in the simulation are summarized in Table 1. PMP mode is employed in our model. Since our proposed scheme is used to recycle the unused bandwidth in UL subframe, the simulation only focuses on the performance of UL transmissions.

CBR is a typical traffic type used to measure the performance of networks in WiMAX research. However, it may not be able to represent the network traffic existing in real life. Moreover, the IEEE 802.16 network aims to serve both data and multimedia applications. Most of the modern streaming videos are encoded by industrial standards (e.g., H.264 or MPEG 4) which generate data in variant rates. In this research, we include VBR traffics to illustrate H.264 and MPEG 4-encoded videos. In our simulation, the traffic models for these streaming videos are based on related research [12], [13], [14]. Additionally, other commonly used VBR traffics such as HTTP and FTP applications are also included in our simulation. The characteristics of traffic types are summarized in Table 2.

In our simulation, each SS serves at least one and up to five connections. Each connection serves one type of traffic which can be mapped to the scheduling classes supported in the IEEE 802.16 standards (i.e., UGS, rtPS, ertPS, nrtPS, and BE). Table 2 enumerates all types of traffics and their corresponding scheduling classes used in our simulation. In particular, all VBR traffics in our simulation are considered as ON/OFF traffics. We fix the mean data rate of each application but make the mean packet size randomly selected from 512 to 1,024 bytes. Thus, the mean packet arrive rate can be determined based on the corresponding mean packet size. As mentioned in our analysis, the size of each packet is modeled as Poisson distribution and the

Fig. 8. Simulation results of *UBR*.

packet arrival rate is modeled as exponential distribution. For example, in order to simulate the network traffics more realistically, the start time of each connection is randomly selected from the 0 to 15th second. Moreover, the real-time connection stops to generate data from the 75th to 100th second. It is for investigating that how good our scheme can achieve when the large amount of unused bandwidth is available. Therefore, the number of active connections (the connections which are transmitting data) may be different during the simulation.

6.2 The Performance Metrics

The simulation used to evaluate the performance of the proposed scheme is based on the three metrics defined as follows:

1. **Throughput gain (TG)**. It represents the percentage of throughput which can be improved by implementing our scheme. The formal definition can be expressed as

$$TG = \frac{T_{\text{recycle}} - T_{\text{no_recycle}}}{T_{\text{no_recycle}}},$$

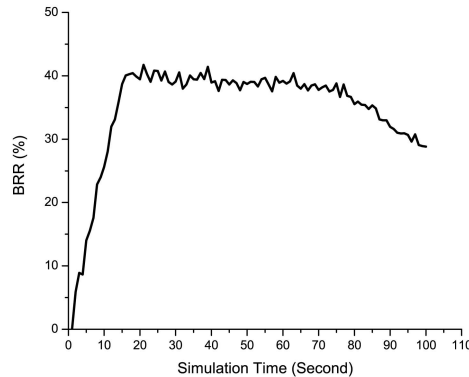
where T_{recycle} and $T_{\text{no_recycle}}$ represent the throughput with and without implementing our scheme, respectively. The higher *TG* achieved shows the higher performance that our scheme can make.

2. **Unused bandwidth rate (UBR)**. It is defined as the percentage of the unused bandwidth occupied in the total granted bandwidth in the system without using bandwidth recycling. It can be defined formally as

$$UBR = \frac{B_{\text{unused_bw}}}{B_{\text{total_bw}}},$$

where $B_{\text{unused_bw}}$ and $B_{\text{total_bw}}$ are the unused bandwidth and total allocated bandwidth, respectively. The *UBR* shows the room which can be improved by our scheme. The higher *UBR* means the more recycling opportunities.

3. **Bandwidth recycling rate (BRR)**. It illustrates the percentage of bandwidth which is recycled from the unused bandwidth. The percentage can be demonstrated formally as

Fig. 9. Simulation results of *BRR*.

$$BRR = \frac{B_{\text{recycled}}}{B_{\text{unused_bw}}},$$

where B_{recycled} is the bandwidth recycled from $B_{\text{unused_bw}}$. *BRR* is considered as the most critical metric since it directly reveals the effectiveness of our scheme.

6.3 Simulation Results

Fig. 8 presents the percentage of the unused bandwidth occupied in our simulation traffic model (i.e., *UBR*). It shows the room of improvement by implementing our scheme. From the simulation results, we can conclude that the average *UBR* is around 38 percent. In the beginning, the *UBR* goes down. It is because each connection still requests bandwidth from the BS. As time goes on, the *UBR* starts to increase when the connection has received the requested bandwidth. After the 75th second of simulation time, *UBR* increases dramatically due to the inactivity of real-time connections. The purpose to have inactive real-time connections is to simulate a network with large amount of unused bandwidth and evaluate the improvement of the proposed scheme in such network status. The evaluation is presented in the later of this section.

The simulation results of recycling rate are presented in Fig. 9. From the figure, we observe that the recycling rate is very close to zero at the beginning of the simulation. It is because that only a few connections transmit data during that time and the traffic load in the system is very light. Therefore, only few connections need to recycle the unused

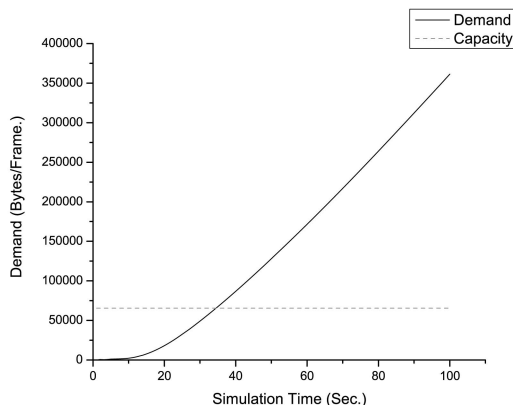
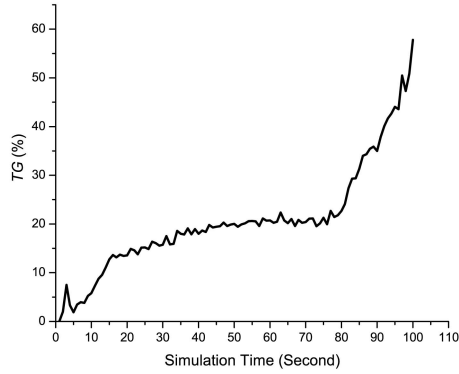


Fig. 10. Total bandwidth demand.

Fig. 11. Simulation results of TG .

bandwidth from others. As time goes on, many active connections join in the network. The available bandwidth may not be able to satisfy the needs of connections. Therefore, there are high probabilities that the CS can recycle the unused bandwidth. It leads a higher BRR .

Fig. 10 shows the total bandwidth demand requested by SSs during the simulation. In the figure, the dashed line indicates the system bandwidth capacity. During the simulation, the BS always allocates the bandwidth to satisfy the demand of real-time connections due to the QoS requirement. Therefore, the amount of bandwidth allocated to non-real-time connections may be shrunk. At the same time, the new non-real-time data are generated. Therefore, the non-real-time data are accumulated in the queue. It is the reason that the demand of bandwidth keeps increasing.

Fig. 11 presents the results of TG calculated from the cases with and without our scheme. In the figure, the TG is very limited at the beginning of the simulation, which is similar to the results of the BRR . It shows Stages 1 and 2 described in Section 5 that there is no significant improvement on our scheme when the network load is light. As the traffic increases, the TG reaches around 15-20 percent. It is worth to note that the TG reaches around 20 percent at the 35th second of the simulation time. It matches the time that the bandwidth demand reaches the system capacity, as shown in Fig. 10. Again, it confirms our early observation (Stages 3 and 4 in Section 5) that the proposed scheme can achieve higher TG when the network is heavily loaded. After the 75th second, the TG increases dramatically. It shows that our scheme can have significant improvement on TG when the large amount of unused bandwidth is available.

We also investigate the delay in the cases with and without our scheme. By implementing our scheme, the average delay is improved by around 19 percent comparing to the delay without using our scheme. It is due to the higher overall system throughput improved by our scheme.

From the simulation results shown above, we can conclude that the proposed scheme cannot only improve the bandwidth utilization and throughput but also decrease the average delay. Moreover, the scheme can have higher performance when the network is heavily loaded. This validates our performance analysis shown in stages 1-4 in Section 5.

Fig. 12 shows the throughput comparison between our scheme and Case with BRs defined in Section 5.6. From the figure, we can obtain that the throughput of Case with BRs

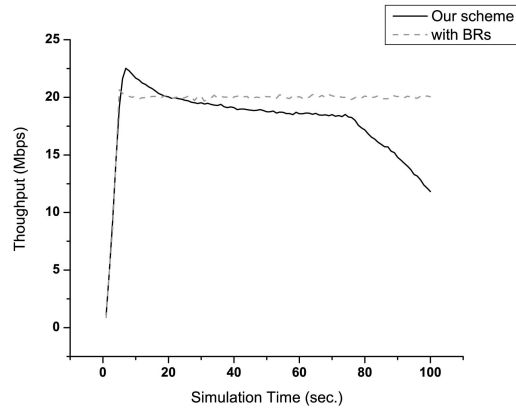


Fig. 12. Comparison with the case with BRs.

can maintain higher throughput than the proposed scheme in most of time but the achievable throughput of our scheme is higher. It is because the SS in the former case always requests bandwidth based on the number of queued data. However, the BS has to reserve sufficient amount of bandwidth for BRs. Therefore, it limits the number of bandwidth for data transmissions. Additionally, this comparison is based on the proposed scheduling algorithm, named Priority-based Scheduling algorithm. The throughput of the proposed scheme is enhanced further by algorithms proposed later in Section 7.

6.4 Theoretical Analysis versus Simulation Results

In this section, we validate the theoretical analysis and simulation results of UBR and RMs coverage. To validate the UBR , we focus on the multimedia traffic specified in Table 2. The simulation model is composed of one BS and one SS. The SS only serves one multimedia traffic specified. The simulation result shows that the UBR is around 35.99 percent. Moreover, the theoretical result calculated by (5) is about 35.29 percent. It is closed to the simulation result.

For validating the coverage of RMs, we employ the typical parameters used in IEEE 802.16 networks in our theoretical analysis. From (20), the theoretical percentage of RMs coverage is from 42 to 58 percent. Additionally, the result from our simulation is 48.7 percent which is within the range of our theoretical result.

To analyze the simulation results more profoundly, we investigate the two factors that the unused bandwidth cannot be recycled: 1) CSs cannot receive RMs sent by their corresponding TSs and 2) CSs do not have data to recycle the unused bandwidth while receiving RMs. According to our simulation results, the probability that a CS fails to recycle the unused bandwidth is around 61.5 percent which includes both factors described above. By doing further investigation, we find that about 51.3 percent of failures is because the CS cannot receive an RM from the corresponding TS. The rest of failures, about 10.2 percent, are caused by no data to be transmitted, while the CS receives an RM. Based on this observation, three scheduling algorithms are proposed in Section 7 to mitigate the affection of these factors for improving the recycling performance.

7 FURTHER ENHANCEMENT

As our investigation, one of the factors causing recycling failures is that the CS does not have data to transmit while receiving an RM. To alleviate this factor, we propose to schedule SSs which have rejected BRs in the last frame because it can ensure that the SS scheduled as CS has data to recycle the unused bandwidth. This scheduling algorithm is called *Rejected Bandwidth Requests First Algorithm* (RBRFA). It is worth to notice that the RBRFA is only suitable to heavily loaded networks with rejected BRs sent from non-real-time connections (i.e., nrtPS or BE). Notice that only rejected BRs sent in the last frame are considered in the RBRFA for scheduling the current frame. The RBRFA is summarized in Algorithm 2.

Algorithm 2. Rejected Bandwidth Requests First Algorithm

Input: T is the set of TSs scheduled on the UL map.

Q_R is the set of SSs which have rejected BRs sent from non-real-time connections in the last frame.

Output: Schedule a CS for each TS in T .

For $i = 1$ to $\|T\|$ **do**

a. $S_t \leftarrow TS_i$.

b. $Q_t \leftarrow Q_R - O_t$.

c. Randomly pick a $SS \in Q_t$ as the corresponding CS of S_t

End For

The BS grants or rejects BRs based on its available resource and scheduling policy. In RBRFA, if the BS grants partially amount of bandwidth requested by a BR, then this BR is also considered as a rejected BR. Similar to Algorithm 1, O_t represents the set of SSs which transmission period overlaps with the TS, S_t , in Q_R . All SSs in Q_t are considered as possible CSs of S_t . A rejected BR shows that the SS must have extra data to be transmitted in the next frame and no bandwidth is allocated for these data. The RBRFA schedules those SSs as CSs on the CL, so the probability to recycle the unused bandwidth while the CS receives the RM can be increased.

The other factor that may affect the performance of bandwidth recycling is the probability of the RM to be received by the CS successfully. To increase this probability, a scheduling algorithm, named as *History-Based scheduling Algorithm* (HBA), is proposed. The HBA is summarized in Algorithm 3. For each TS, the BS maintains a list, called *Black List* (BL). The basic CID of a CS is recorded in the BL of the TS if this CS cannot receive RMs sent from the TS. According to our protocol, the CS will transmit data or pad the rest of transmission interval if an RM is received. The BS considers that a CS cannot receive the RM from its corresponding TS if the BS does not receive either data or padding information from the CS. When the BS schedules the CS of each TS in future frames, the BS only schedules an SS which is not on the BL of the TS as the CS. After collecting enough history, the BL of each TS should contain the basic CID of all SSs which cannot receive the RM sent from the TS. By eliminating those SS, the BS should have high probability to schedule a CS which can receive the RM successfully. Therefore, HBA can increase the probability of scheduling an SS which is able to receive the RM as the CS.

Algorithm 3. History-Based Scheduling Algorithm

Input: T is the set of TSs scheduled on the UL map.

Q is the set of SSs running non-real time applications

BL is the set of black lists of TSs.

Output: Schedule a CS for each TS in T .

For $i = 1$ to $\|T\|$ **do**

a. $S_t \leftarrow TS_i$.

b. $Q_t \leftarrow Q - O_t - BL_i$

c. Randomly pick a $SS \in Q_t$ as the corresponding CS of S_t

d. **IF** the scheduled CS did not transmit data or SBV

Then put this CS in the BL_i

End For

To support the mobility defined in IEEE 802.16e standard, the BL of each TS should be updated periodically. Moreover, the BS changes the UL burst profile of the SS when it cannot listen to the SS clearly. There are two possible reasons which may make the BS receive signals unclearly: 1) the SS has moved to another location and 2) the background noise is strong enough to interfere the data transmissions. Since those two factors may also affect the recipient of RMs; therefore, the BL containing this SS should be updated as well.

The two algorithms described above focus on mitigating each factor that may cause the failure of recycling. The RBRFA increases the probability that the CS has data to transmit while receiving the RM. The HBA increases the probability that the CS receives the RM. However, none of them can alleviate both factors at the same time. By taking the advantages of both RBRFA and HBA, an algorithm called *Hybrid Scheduling Algorithm* (HSA) is proposed. HSA can increase not only the probability of CSs to transmit data while receiving the RM, but also the probability of CSs to receive the RM. The detail of HSA is summarized in Algorithm 4.

Algorithm 4. Hybrid Scheduling Algorithm

Input: T is the set of TSs scheduled on the UL map.

Q_R is the set of SSs which have rejected BRs sent for non-real time applications.

BL is the set of black lists of TSs.

Output: Schedule a CS for each TS in T .

For $i = 1$ to $\|T\|$ **do**

a. $S_t \leftarrow TS_i$.

b. $Q_t \leftarrow Q_R - O_t - BL_i$

c. Randomly pick a $SS \in Q_t$ as the corresponding CS of S_t

d. **IF** the scheduled CS did not transmit data or SBV

Then put this CS in the BL_i

End For

When the BS schedules the CS for each TS, only the SSs with rejected BRs are considered. As mentioned before, it can increase the probability of CSs to transmit data while receiving the RM. Moreover, the BS maintains a BL for each TS. It can screen out the SSs which cannot receive the RM so that those SS cannot be scheduled as the CSs. The probability of receiving RMs can be increased. Again, the BL of each TS should be updated periodically or when the UL burst profile of the SS has been changed. By considering those two advantages, HSA is expected to achieve higher *TG* and *BBR* comparing to RBRFA and HBA.

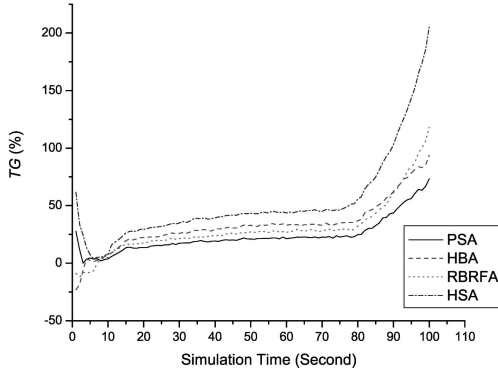


Fig. 13. Simulation results of TG among all scheduling algorithms.

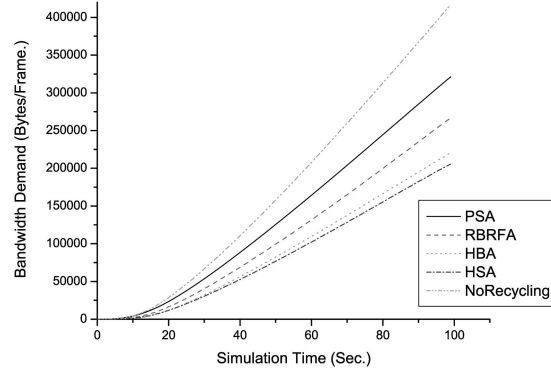


Fig. 15. Simulation results of bandwidth demand.

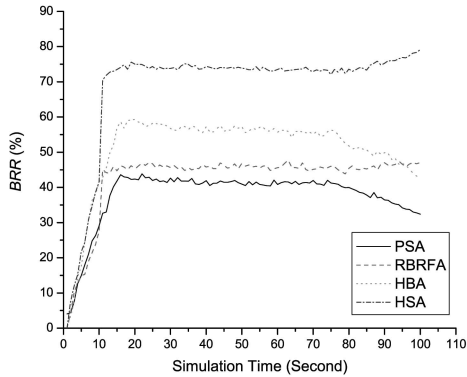


Fig. 14. Simulation results of BRR among all scheduling algorithms.

8 SIMULATION RESULTS OF ENHANCEMENT

The simulation model for evaluating these scheduling algorithms is same as the model presented in Section 6. The BS is located at the center of a geographical area. There are 50 SSs uniformly distributed in the service coverage of BS. Each SS serves at least one and up to five connections. The simulation results of TG are shown in Fig. 13. Before the 15th second of simulation time, the TG may be negative. It means that the throughput without recycling is higher than the throughput with recycling. It is because the applications of each SS start to generate data randomly in the first 15 seconds of simulation time. As described before, the PSA shown as Algorithm 1 can achieve averagely 20 percent of throughput. The RBRFA can further improve the throughput to 26 percent because of increasing the chance of transmitting data, while the CS receives the RM. Moreover, the HBA can have a greater improvement on TG to 30 percent. It shows that the factor of missing RMs causes more failures of recycling than the factor of no data transmissions, while the CS receives the RM does. This result consists with our observation in Section 6 that the probability of missing RMs is higher than the probability that the CS cannot recycle the unused bandwidth due to the lack of data to be transmitted. Moreover, HSA achieves the best performance on TG (averagely 45 percent improvement) since it combines both advantages of HBA and RBRFA.

The comparison of BRR is shown in Fig. 14. The results consist with the results of TG shown above. The HSA has the highest BRR . Moreover, the HBA achieves the higher BRR than the RFA does. Additionally, it is worth noting that the BRR of the RRFA cannot be more than 50 percent even

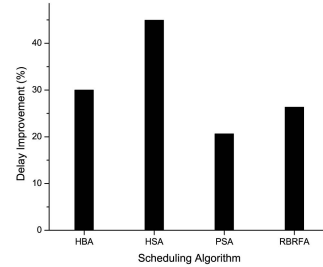


Fig. 16. Simulation results of delay improvement.

when the network is fully loaded. It is because, based on our investigation in Section 6, there is only 48.7 percent of probability that a CS can receive an RM successfully.

The comparison of the total bandwidth demand is shown in Fig. 15. From the figure, the increasing speed of bandwidth demand from low to high is HSA, HBA, RBRFA, PSA, and No Recycling. This result matches the result of TG . It is because that there are fewer data accumulated in the queue when the TG is higher. It leads to less bandwidth demand.

Due to the improvement of throughput, the average delay is also improved. The summary of delay improvement is shown in Fig. 16. Similar to the simulation results of TG and BRR , the HSA has the best improvement on delay due to the highest throughput it achieves.

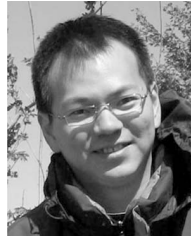
9 CONCLUSIONS

Variable bit rate applications generate data in variant rates. It is very challenge for SSs to predict the amount of arriving data precisely. Although the existing method allows the SS to adjust the reserved bandwidth via bandwidth requests in each frame, it cannot avoid the risk of degrading the QoS requirements. Moreover, the unused bandwidth occurs in the current frame cannot be utilized by the existing bandwidth adjustment since the adjusted amount of bandwidth can be applied as early as in the next coming frame. Our research does not change the existing bandwidth reservation to ensure that the same QoS guaranteed services are provided. We proposed *bandwidth recycling* to recycle the unused bandwidth once it occurs. It allows the BS to schedule a complementary station for each transmission stations. Each complementary station monitors the entire UL transmission interval of its corresponding TS and standby for any opportunities to recycle the unused bandwidth. Besides the naive priority-based scheduling algorithm, three addi-

tional algorithms have been proposed to improve the recycling effectiveness. Our mathematical and simulation results confirm that our scheme cannot only improve the throughput, but also reduce the delay with negligible overhead and without degrading the QoS requirements.

REFERENCES

- [1] IEEE 802.16 WG, *IEEE Standard for Local and Metropolitan Area Network Part 16: Air Interface for Fixed Broadband Wireless Access Systems*, IEEE Std 802.16-2004, IEEE, pp. 1-857.
- [2] IEEE 802.16 WG, *IEEE Standard for Local and Metropolitan Area Networks Part 16: Air Interface for Fixed and Mobile Broadband Wireless Access Systems, Amendment 2*, IEEE, Dec. 2005.
- [3] J. He, K. Yang, and K. Guild, "A Dynamic Bandwidth Reservation Scheme for Hybrid IEEE 802.16 Wireless Networks," *Proc. IEEE Int'l Conf. Comm. (ICC '08)*, pp. 2571-2575.
- [4] K. Gakhar, M. Achir, and A. Gravey, "Dynamic Resource Reservation in IEEE 802.16 Broadband Wireless Networks," *Proc. IEEE Int'l Workshop Quality of Service (IWQoS)*, pp. 140-148, 2006.
- [5] J. Tao, F. Liu, Z. Zeng, and Z. Lin, "Throughput Enhancement in WiMax Mesh Networks Using Concurrent Transmission," *Proc. IEEE Int'l Conf. Wireless Comm., Networking and Mobile Computing*, pp. 871-874, 2005.
- [6] X. Bai, A. Shami, and Y. Ye, "Robust QoS Control for Single Carrier PMP Mode IEEE 802.16 Systems," *IEEE Trans. Mobile Computing*, vol. 7, no. 4, pp. 416-429, Apr. 2008.
- [7] E.-C. Park, H. Kim, J.-Y. Kim, and H.-S. Kim, "Dynamic Bandwidth Request-Allocation Algorithm for Real-Time Services in IEEE 802.16 Broadband Wireless Access Networks," *Proc. IEEE INFOCOM*, pp. 852-860, 2008.
- [8] T.G. Robertazzi, *Computer Networks and Systems: Theory and Performance Evaluation*. Springer-Verlag, 1990.
- [9] K. Gakhar, M. Achir, and A. Gravey, "How Many Traffic Classes Do We Need In WiMAX?" *Proc. Wireless Comm. and Networking Conf. (WCNC)*, pp. 3703-3708, 2007.
- [10] G. Iazeolla, P. Kritzing, and P. Pileggi, "Modelling Quality of Service in IEEE 802.16 Networks," *Proc. IEEE Conf. Software, Telecomm. and Computer Networks (SoftCOM)*, pp. 130-134, 2008.
- [11] Qualnet, http://www.scalable-networks.com/products/developer/new_in_45.php, 2010.
- [12] F.H.P. Fitzek and M. Reisslein, "MPEG-4 and H.263 Video Traces for Network Performance Evaluation," *IEEE Network*, vol. 15, no. 6, pp. 40-54, Nov./Dec. 2001.
- [13] P. Seeling, M. Reisslein, and B. Kulapala, "Network Performance Evaluation Using Frame Size and Quality Traces of Single-Layer and Two-Layer Video: A Tutorial," *IEEE Comm. Surveys and Tutorials*, vol. 6, no. 2, pp. 58-78, July-Sept. 2004.
- [14] G. Van der Auwera, P.T. David, and M. Reisslein, "Traffic and Quality Characterization of Single-Layer Video Streams Encoded with H.264/AVC Advanced Video Coding Standard and Scalable Video Coding Extension," *IEEE Trans. Broadcasting*, vol. 54, no. 3, pp. 698-718, Sept. 2008.



Network Innovation Labs at Yokosuka, Japan, in 2009.



Microsystems and IT Professional. He is a senior member of the IEEE.

David Chuck received the BS degree in applied mathematics from Tatung University, Taiwan, in 2004, and the MS degree in computer science from the Illinois Institute of Technology in 2007. He is currently working toward the PhD degree in computer engineering at Iowa State University. His research interests include wireless networking, network virtualization, resource allocation, and game-theoretical study of networks. He also worked as a research intern in the NTT

J. Morris Chang received the PhD degree from North Carolina State University. He is an associate professor at Iowa State University. His industrial experience includes Texas Instruments, and AT& T Bell Labs. He received the University Excellence in Teaching Award at Illinois Institute of Technology in 1999. His research interests include wireless networks and computer systems. Currently, he is an editor of the *Journal of Microprocessors and*

► For more information on this or any other computing topic, please visit our Digital Library at www.computer.org/publications/dlib.