2012

Bandwidth and Power Management in Broadband Wireless Networks

David Haoen Chuck
Iowa State University

Bandwidth and power management in broadband wireless networks

by

David Haoen Chuck

A dissertation submitted to the graduate faculty

in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Major: Computer Engineering

Program of Study Committee:

Morris Chang, Major Professor

Yong Guan

Ahmed Kamal

Lu Ruan

Lei Ying

Iowa State University

Ames, Iowa

2012

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Acknowledgements

I would like to take this opportunity to thank those who helped and supported me in various aspects of conducting research and writing this thesis.

First, I would like to thank Dr. Morris Chang for his guidance throughout this research and writing of this dissertation. His experience and encouragement have often inspired me to conduct innovation research and complete my graduate education. I would also like to thank my committee members for their suggestions and contributions to this work: Dr. Ahmed E. Kamal, Dr. Lei Ying, Dr. Yong Guan and Dr. Lu Ruan. Additionally, I would like to thank Mr. Noriyuki Takahashi and Dr. Girija Narlikar for internship opportunities in their labs. These experiences enrich my view of conducting research.

At last, I would like to thank my family and Meng-hsien Lin for their endless love and support in both life and study.

Abstract

Bandwidth and power are considered as two important resources in wireless networks. Therefore, how to management these resources becomes a critical issue. In this thesis, we investigate this issue majorally in IEEE 802.16 networks. We first perform performance analysis on two bandwidth request mechanisms defined in IEEE 802.16 networks. We also propose two practical performance objectives. Based on the analysis, we design two scheduling algorithm to achieve the objectives.

Due to the characteristics of popular variable bit rate (VBR) traffic, it is very difficult for subscriber stations (SSs) to make appropriate bandwidth reservation. Therefore, the bandwidth may not be utilized all the time. We propose a new protocol, named bandwidth recycling, to utilized unused bandwidth. Our simulation shows that the proposed scheme can improve system utilization averagely by 40%.

We also propose a more aggressive solution to reduce the gap between bandwidth reservation and real usage. We first design a centralized approach by linear programming to obtain the optimal solution. Further, we design a fully distributed scheme based on game theory, named bandwidth reservation (BR) game. Due to different quality of service (QoS) requirements, we customize the utility function for each scheduling class. Our numerical and simulation show that the gap between BR game and optimal solution is limited.

Due to the advantage of dynamical fractional frequency reuse (DFFR), the base station (BS) can dynamically adjust transmission power on each frequency partition. We emphasis on power allocation issue in DFFR to achieve most ecomicical data transmission. We first formulate the problem by integer linear programming (ILP). Due to high computation complexity, we further design a greedy algorithm. Our simulation shows that the results of the greedy algorithm is very close to the ILP results.

CHAPTER 1. General Introduction

1.1 Introduction

Wireless broadband networks have received significant attraction recently. With the increase of traffic demand, how to efficiently utilize bandwidth becomes a critical issue in wireless network. In order to support high data rate transmission, the equipment consumes more power. Thus, power conservation is also important in wireless networks. In this thesis, we investigate these two critical important topics in wireless networks. We mainly focus on the technologies based on IEEE 802.16 standard. The Worldwide Interoperability for Microwave Access (WiMAX), based on this family of standard, 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) layer of WiMAX have been specified in the IEEE 802.16 standard. Many advanced communication technologies such as OFDM/OFDMA and MIMO are employed in this family of standard. Supported by these modern technologies, WiMAX is able to provide a large service coverage, a high speed data rate and quality of service (QoS) guaranteed services. Because of these features, WiMAX is considered to be a promising alternative for last mile broadband wireless access (BWA). In this section, we first introduce the background information of IEEE 802.16 networks and then a brief summary of each project attached to this document is present in each section.

1.2 Background Information

One of the fundamental features in IEEE 802.16 networks is to provide QoS guaranteed services. Radio resource reservation is employed in the IEEE 802.16 standard to achieve

this feature. In order to serve wide variety of applications, all traffics from upper layer are mapped into connections. Each connection is classified into one of five scheduling classes depending on the QoS requirements of applications. The detail of these scheduling classes defined in the IEEE 802.16 standard are presented below:

- Unsolicited Grant Service (UGS) :

  This scheduling class has the highest priority among all scheduling classes. It is designed to support real-time data stream consisting of fixed-size packets issued at periodic intervals such as T1/E1 and Voice over IP (VoIP) without silence compression. Because of designing for traffic with fixed-size packets at periodic intervals, the minimum reserved traffic rate should be equal to the maximum sustained traffic rate. It is worth to note that the amount of bandwidth allocated to a UGS connection is unsolicited and no bandwidth requests are allowed for UGS connections.

- Real Time Polling Service (rtPS) :

  It is designed to support the real-time stream data with variable size packets issued at periodic intervals, e.g. MPEG video. To ensure the QoS, the BS periodically gives unicast polling opportunities to the SS. Thus, the SS requests bandwidth without contending with other SSs. The bandwidth is allocated after the bandwidth request is transmitted. Therefore, rtPS involves an additional delay in the bandwidth request-allocation process.

- Extended Real Time Polling Service (ertPS) :

  This class is defined in IEEE 802.16e-2005 and designed to support real-time traffics with variable data rate and requiring guaranteed data rate and delay such as VoIP with silence compression. This is basically identical to UGS. However, the difference between ertPS and UGS is that ertPS can change the amount of allocated bandwidth depending on the traffic characteristics. In order to ensure the QoS, ertPS is allowed to use both unicast polling and contention resolution to request bandwidth.

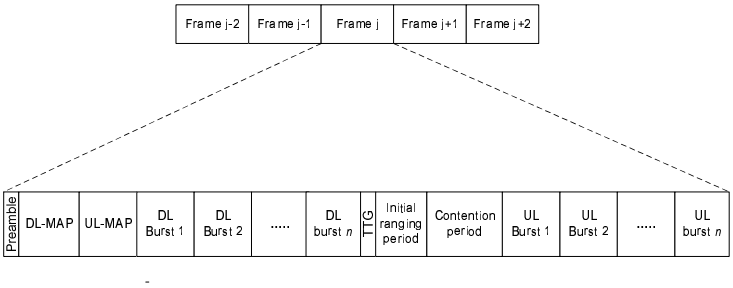


Figure 1.1 Frame Structure

- Non-real Time Polling Service (nrtPS) :

  This class is designed to serve delay-tolerant data streams, such as FTP, which requires variable-size data grants at a minimum guaranteed rate. Thus, the minimal reserved rate can not be set to zero. The nrtPS is allowed to use both unicast polling and contention opportunities to request bandwidth.

- Best Effort (BE) :

  This class is designed to serve data streams without requiring a minimum guaranteed rate such as web browsing. This is allowed to use the same bandwidth request mechanisms as nrtPS. This class has the lowest service priority.

Based on the scheduling classes, a request/grant bandwidth allocation mechanism based on each connection is specified in the IEEE 802.16 standard. Each subscriber station (SS) requests the required bandwidth from the base station (BS) followed by the QoS requirements of each connection via bandwidth requests. After receiving a request, the BS make scheduling decision to determine the amount of bandwidth allocated to each SS. The SS receives exclusive privilege to utilize the allocated bandwidth. All bandwidth reservation are expressed in a MAC frame.

A MAC frame in IEEE 802.16 networks comprises two subframes: downlink(DL) subframe and uplink(UL) subframe. The DL subframe is responsible for the transmissions from base station (BS) to subscriber stations (SSs). On the other hand, the UL subframe is for

the transmissions in the opposite direction. The detail frame structure is shown in Fig. 1.1. At the beginning of DL subframe, the BS broadcast two messages: DL-MAP and UL-MAP to all served SSs. Those two messages contain the information for data transmissions between BS and each SS. Thus, each SS can prepare its data transmissions after receiving these two messages.

1.3 A Comprehensive Analysis of Bandwidth Request Mechanisms in IEEE 802.16 Networks

As mentioned earlier, a request/grant bandwidth allocation is introduced in IEEE 802.16 networks. The SS can request the required bandwidth via bandwidth requests (BRs). There are two BR mechanisms for the SS to transmit BRs: unicast polling and contention resolution. The former relies on the BS scheduling the amount of bandwidth on the top of the existing bandwidth reservation of the SS. Since the SS has the privilege to utilize allocate bandwidth, the unicast polling can ensure the SS to have opportunities to make BRs. Consequently, the delay of BR transmissions can be controlled. However, due to the privilege, the allocated bandwidth may be wasted if the SS does not transmit any BRs. The BS schedules the amount of bandwidth to a group of SSs for making BRs in contention resolution. Each SS in the group should contend with each other for granting a transmission opportunity for making BRs. Thus, the bandwidth may not be wasted because some of SSs do not transmit BRs. However, due to the nature of contention, the SS may not able to have opportunities for BR transmissions. Thus, the transmission delay may not be guaranteed.

As specified in the IEEE 802.16 standard, the connection in all scheduling classes except UGS are allowed to be scheduled unicast polling opportunities. However, only ertPS, nrtPS and BE connections allow to grant BR transmission opportunities via contention resolution. Since each BR mechanism has its own advantages and disadvantages, it becomes a very critical challenge for the BS to schedule the BR mechanisms to the connections which are are allowed to use both BR mechanisms for BR transmissions such that the the performance

objectives of the network are satisfied. In this document, we first analyze comprehensively of both BR mechanisms. Additionally, we propose two potential performance objectives: 1) how to maximize the throughput while satisfying the fixed delay requirement. 2) how to minimize the delay with the minimum required throughput. Based on the results of analysis, we propose two algorithms to achieve these objectives. Our numerical and simulation have confirmed that the scheduling algorithms can always have better performance by scheduling one of the BR mechanisms. Our research results have been published at *IEEE Transactions on Vehicular Technology*.

1.4 Bandwidth Recycling in IEEE 802.16 Networks

The reservation scheme is employed in the IEEE 802.16 networks. The SS has to determine the amount of bandwidth to reserve and no one can utilize the reserved bandwidth except the reserving SS. Variable bit rate (VBR) applications generate traffic in a unsettle rate. Because of this nature, it is very challenging for the SS to make optimal bandwidth reservation to serve VBR applications. Consequently, the reserved bandwidth may be wasted if the reserving SS has no data to transmit. Although the SS can adjust the bandwidth reservation via BRs, however, the adjusted bandwidth reservation is enforced as early as in the next coming frame. There is no way to utilize the unused bandwidth in the current frame. Additionally, the BR adjusts the bandwidth reservation permanently. It may expose the QoS of applications in danger since the SS may not be able to receive the desired amount of bandwidth. To alleviate this problem, we proposed a scheme, called *Bandwidth Recycling*, to utilize the unused bandwidth in the currently frame while providing QoS guaranteed services.

The main idea of bandwidth recycling is to schedule a backup SS to recycling the unused bandwidth for each SS scheduled on the UL-MAP. In our scheme, the reserving bandwidth transmits a message, called *releasing message (RM)*, to the backup SS when the unused bandwidth is available. The backup SS starts to utilize the unused bandwidth after receiving

the RM. There are two reasons that the bandwidth may not be recycled successfully: 1) the backup SS does not receive the RM. 2) the backup SS does not have data to recycle the unused bandwidth while receiving the RM. To alleviate these two reasons, we proposed three scheduling algorithms. Based on our simulation, the proposed scheme can improve the average system throughput by 40%. This research work has been published in *IEEE Transactions on Mobile Computing*.

1.5 Design and Analysis of Bandwidth Reservation Game in IEEE 802.16 Networks

As described in Section 1.4, it is very challenging for the SS to make the optimal bandwidth reservation to serve VBR applications such that the QoS requirements are satisfied. Consequently, it is a critical issue to help the SS make the appropriate bandwidth reservation. In additional to satisfy QoS requirements, the SS which bandwidth reservation mismatches the bandwidth demand may degrade the performance of the entire network. The degree of this performance degradation is related to the current bandwidth demand of all SSs in the network. For example, the SS may cause a very limited degradation in the network with a light load. Moreover, the data latency may be reduced when the SS requests more bandwidth in a light loaded network. However, the network performance may have a huge degradation when the network is fully loaded. Consequently, the optimal bandwidth reservation relates to not only the QoS requirements of applications but also the bandwidth demand of the entire network. In our work, we propose a game theoretical framework to assist the SS to reach the optimal bandwidth reservation.

In our game, the player are SSs. Each SS focuses on maximizing the utility profit based on the utility function. The utility function comprises two indexes: Satisfaction (SI) Index and Penalty Index (PI). Since the most fundamental duty of the SS is to ensure that the QoS requirements of applications are satisfied, the SI is the index to represent the QoS satisfaction of the application corresponding to the amount of reserved bandwidth. At the

same time, the SS may cause network performance degradation with the amount of reserved bandwidth. The PI, on the other hand, is designed to reflect the network performance degradation caused by the SS corresponding to the amount of reserved bandwidth. The utility profit is defined as $(SI - PI)$. Each SS is try to reserve the amount of bandwidth such that the utility profit is maximized. In our work, we have proved that the existence and uniqueness of Nash equilibrium. Additionally, our numerical results show that the SS can request more bandwidth to reduce the data latency in the network with a light load and satisfy the QoS requirements in a heavily loaded network. This research has been submitted to *IEEE Transactions on Mobile Computing.*

1.6 Economical Data Transmission in Dynamical Fractional Frequency Reuse

Fractional frequency reuse (FFR) is proposed to improve the spectrum utilization such that higher transmission rate is achieved. Unlike the conventional frequency assignment, FFR allows each BS to utilize all possible frequency sections but allocate different levels of transmission power to avoid interference. Recently, dynamical fraction frequency reuse (DFFR) is proposed. It allows BS to dynamically adjust the transmission power of each frequency section. Consequently, power allocation in both FFR and DFFR may directly affect not only the system performance in an individual cell but also the surrounding cells. An efficient power allocation mechanism is desired to manage the power allocation between BSs. Furthermore, providing QoS guaranteed services is one of the fundamental feature in next generation networks. In our research, we study the power allocation problem to help the BS manage the transmission power while maintaining QoS guaranteed services. We propose a joint optimization of system throughput as well as power consumption. We first model the problem by integer linear program. Due to high computation complexity, we further design a heuristic algorithm for practical implementation. The performance evaluation results show that the heuristic algorithm can achieve almost optimal solution.

This thesis is organized as follows. In Chapter 2, we first present the analysis of bandwidth recycling mechanism. The bandwidth recycling is placed in Chapter 4. The more aggressive solution for bandwidth allocation (i.e., bandwidth reservation game) is presented in Chapter 4. Finally, we discuss economical data transmission in DFFR in Chapter 5. The conclusion and future work is given in Chapter 6.

CHAPTER 2. A Comprehensive Analysis of Bandwidth Request Mechanisms in IEEE 802.16 Networks

A paper to be published in IEEE Transactions on Vehicular Technology

Volume: 59, Issue: 4 Page(s): 2046 - 2056

David Chuck and J. Morris Chang

Abstract

IEEE 802.16 standard is considered as one of the most promising technologies. Bandwidth reservation is employed to provide quality of service (QoS) guaranteeing services. A request/grant scheme is defined in the IEEE 802.16 standard. There are two types of bandwidth request (BR) mechanisms, *unicast polling* and *contention resolution*, defined in the standard. As specified, connections belonging to scheduling classes of ertPS, nrtPS and BE have options to make BRs via both mechanisms depending on the scheduling decision made by the base station (BS). However, most research works only assume one of them is available and do not take both of them into account. A comprehensive study of both mechanisms is critical for the BS to make an appropriate decision for those connections to achieve better system performance. To the best of our knowledge, this is the first attempt to analyze this issue. There are two major contributions presented in this paper. First, a comprehensive study of both BR mechanisms in terms of bandwidth utilization and delay is provided. Additionally, we propose two practical performance objectives: when the expected delay or target bandwidth utilization is given, how does the BS to make scheduling decision such that the performance of the other metric (either delay or bandwidth utilization) is optimized? As our second contribution, we proposed two scheduling algorithms to

find the combination of both mechanisms to meet our objectives. The simulation results show that our scheduling algorithms can always help the BS make scheduling decision to reach better system performance.

2.1 Introduction

The IEEE 802.16 standards (e.g., 802.16-2004 (1)) are considered as one of critical broadband wireless access (BWA) technologies in the forth generation (4G) networks. 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 OFDM/OFDMA and MIMO are embraced in the standards. Supported by these modern technologies, WiMAX is able to provide a large service coverage, a high speed data rate and quality of service (QoS) guaranteeing services. Because of these features, WiMAX is considered to be a promising alternative for last mile BWA.

In order to provide the QoS guaranteeing services, bandwidth reservation is adopted in the WiMAX network. A request/grant bandwidth allocation is employed for reserving bandwidth. The subscriber station (SS) is required to reserve the sufficient amount of bandwidth from the base station (BS) before any data transmissions. The amount of reserved bandwidth can be reserved or adjusted by the SS via sending bandwidth requests (BRs). There are two type of bandwidth requests specified in the IEEE 802.16 standard: *unicast polling* and *contention resolution*. In unicast polling, the BS allocates a small piece of bandwidth to the target SS. This small piece of bandwidth is on the top of reserved bandwidth and contains one or more transmission opportunities (TxOPs) depending on the scheduling policy that the BS enforces. These TxOPs are called *unicast polling TxOPs* in this paper. The target SS can use these TxOPs to send BRs. Moreover, for simplicity, we assume that the unicast polling TxOP is only used for transmitting a BR. Contention

resolution, on the other hand, requires that each SS contends a TxOP independently to transmit a BR. The BS schedules an amount of bandwidth, divided into several TxOPs, for a group of SSs to send BRs. These TxOPs are called *contention TxOPs*. If the attempt of contention is failed, then the SS enters into the back-off procedure to prepare next attempt until reaching the maximum number of attempts.

Each type of BR mechanisms (i.e. unicast polling or contention resolution) has its own advantages and disadvantages. In unicast polling, the unicast polling TxOPs are allocated exclusively for the target SS. It guarantees that this SS has opportunities to make BRs successfully. Therefore, the delay of the SS to transmit a BR can be bounded within a certain range. However, because of the exclusive usage, the allocated unicast polling TxOPs are wasted if the target SS does not make BRs. This may reduce the bandwidth utilization of the system. In contention resolution, on the other hand, the allocated bandwidth is shared by a group of SSs. The SS contends with each other in order to get a contention TxOP for the BR. In the contention resolution, each SS contends for a contention TxOP actively. Therefore, the SS performs the contention procedure only if the SS wants to transmit a BR. It may lead to higher bandwidth utilization. However, Each SS cannot be guaranteed to have contention TxOPs for sending BRs. Thus, the delay to request bandwidth can not be ensured.

Support for QoS is a fundamental part of the IEEE 802.16 MAC-layer design. When the service data unit arrives in the IEEE 802.16 MAC layer, the classification process is performed. The classification process is the process which maps the service data unit to the appropriate scheduling class based on the QoS constraints of the service data unit. As specified in the IEEE 802.16 standard, only connections belonging to three scheduling classes (i.e. extended real time polling service (ertPS), non-real time polling service (nrtPS) and best effort (BE)) are allowed to have options to choose between unicast polling and contention resolution for make BRs. Because of the features of each BR mechanism, A scheduling decision made by the BS for the connections in these scheduling classes to

transmit BRs may affect the overall bandwidth utilization and delay. For example, unicast polling may result low bandwidth utilization when the probability of a SS to make a BR is low. Similarly, the contention resolution may leads to a large number of collisions when the probability that a SS makes BRs is high. The motivation of this research is "how does the BS schedule those two types of BR mechanisms to serve the SS while maintaining good system performance?". An appropriate decision made by the BS is needed in order to achieve the desired performance objectives. Thus, the impact of this research is to help the BS make scheduling decisions between two types of BR mechanism specified in the standard in order to meet our performance objectives.

There are two proposed performance objectives considered in this paper: *1)* Maximizing the bandwidth utilization while satisfying the desired delay. *2)* Minimizing the delay while maintaining the target bandwidth utilization. To achieve the performance objectives respectively, two scheduling algorithms are proposed in Section 2.5: $MAX-U$ (for the first objective) and $MIN-D$ (for the second objective). Many research works related to those two BRs mechanisms are only focused on the optimization of one type of BRs mechanisms based on the assumption that only one type of BR mechanisms is available to be used. A comprehensive study considering both mechanisms is desired for the BS to schedule the connections which are allowed to send BRs via both mechanisms. In this paper, we provide mathematical analysis for both BR mechanisms. Based on the analysis, we proposed two scheduling algorithms for performance objectives to help the BS make an appropriate scheduling decision such that the system can have better performance.

The rest of paper is organized as follows. An overview of IEEE 802.16 and the related work are provided in Section 2.2 and Section 2.3, respectively. Our mathematical analysis of both BRs mechanisms is given in Section 2.4. In Section 2.5, we introduce the objectives and proposed scheduling algorithms. Section 2.6 presents the simulation and Section 2.7 concludes our discussion.

2.2 Overview of IEEE 802.16

A IEEE 802.16 network is composed by a number of SSs and at least one BS. There are two operational modes, point-to-multipoint (PMP) and mesh, defined in the IEEE 802.16 standard. This paper is focused on the PMP mode which defines that transmissions are only allowed between the BS and SSs. All transmissions can be classified into downlink (DL) and uplink (UL) transmission based on the direction of transmissions. The DL transmission is defined as the transmission from the BS to a SS. Conversely, the UL transmission is the transmission in the opposite direction. According to the IEEE 802.16 standard, the BS is responsible for scheduling both UL and DL transmissions. All scheduling behavior is expressed in a MAC frame.

The structure of a MAC frame defined in the IEEE 802.16 standard can be divided into UL subframe and DL subframe. The UL subframe is for UL transmissions. Similarly, the DL subframe is for DL transmissions. In a IEEE 802.16 network, all SSs should be coordinated by the BS. All coordinating information including burst profiles and offsets is resided in the DL and UL maps, which are broadcasted at the beginning of the MAC frame.

The IEEE 802.16 network is connection-oriented. It requires SSs to establish connections with the BS before any data transmissions. In order to support wide variety of applications, the IEEE 802.16 standard classifies all traffics into five scheduling classes based on the different QoS requirements: Unsolicited Grant Service (UGS), Real Time Polling Service (rtPS), Extended Real Time Polling Service (ertPS), Non-real Time Polling Service (nrtPS) and Best Effort (BE).

The mechanism to make BRs for each scheduling class has been specified in the IEEE 802.16 standard. For example, a fixed amount of bandwidth is given to UGS connections and BRs are prohibited to be made for this type of connections. All connections in other scheduling classes (i.e. rtPS, ertPS, nrtPS and BE) are allowed to make BRs via unicast polling opportunities. However, ertPS, nrtPS and BE connections are the only connections which are allowed to request bandwidth via contention resolution.

The operation procedure of unicast polling defined in the IEEE 802.16 standard is straight forward. The BS allocates an extra piece of bandwidth to the target SS. This extra piece of bandwidth can be considered as one or more unicast polling TxOPs. The target SS makes bandwidth requests by utilizing these TxOPs. Since these TxOPs are exclusively allocated to this particular SS, it can ensure that this SS has opportunities to request bandwidth if needed. However, the drawback is that these TxOPs are wasted if this SS does not make BRs.

The contention resolution, on the other hand, is not TxOPs-guaranteed. It means that the SS may not have opportunities to transmit BRs due to the failures of contention. The BS schedules a few contention TxOPs for a group of SSs. Each SS within this group is required to contend for a contention TxOP with each other in order to transmit a BR. Note that each contention TxOP can only carry one BR. If the SS fails in the contention procedure, it enters into the back-off procedure for preparing the next attempt. In this paper, the binary exponential back-off (BEB) algorithm (11) is employed as the back-off procedure. The initial back-off window size and the maximum back-off window size are controlled by the BS and specified in the UL map. The value of contention window size is represented as a power-of-two value. For example, a value of 4 indicates that the contention window size is 16.

The operation procedure of contention resolution is summarized as Fig.2.1. When a SS tends to contend a TxOP, it selects a random number from 0 to $W - 1$, where W is the current back-off window size. This random number is called back-off counter and indicates the number of contention TxOPs that the SS shall defer before transmitting. The number of contention TxOPs is determined by the BS and may be different in each frame. If the back-off counter does not reach zero within a contention period. Its countdown should be frozen at the end of the contention period and resume at the beginning of the next coming contention period.

When the back-off counter reaches zero, the SS attempts to send a BR. It is possible

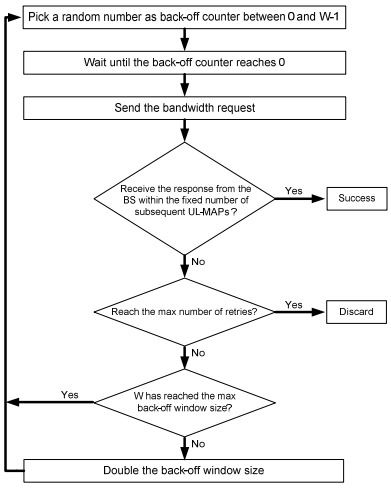


Figure 2.1 The operation of contention resolution

that there are more than one SS which back-off counter reaches zero at the same time. It means that there are more than one SSs trying to send a BR in the same TxOP. In this case, collision happens. Since it is not practically possible for SSs to sense the UL channel to detect a collision, the SS can only know the success of BR transmission if it receives a response from the BS in the form of bandwidth grant within a fixed number of subsequent UL map messages. If the SS fails to receives the response, it considers that the BR was not delivered successfully. The SS shall double its back-off window size if the current contention window size is smaller than the maximum back-off window size which is controlled by the BS. The SS selects a fresh random number from 0 to $W' - 1$, where the W' indicates the new back-off window size, and repeat the deferring procedure described above. The SS can attempt to transmit BRs until the maximum number of retries is reached.

2.3 Related Work

Many research works related to unicast polling and contention resolution have been proposed in the literature. In (3), an adaptive polling scheme for ON/OFF traffic was proposed to improve the bandwidth utilization for unicast polling. During ON periods, polling intervals are fixed and short, while during OFF periods polling intervals are lengthened ex-

ponentially. Therefore, adaptive polling reduces the signaling overhead without significantly compromising delay performance. A Markov chain model for unicast polling is proposed in (4). The authors proposed the Markov chain analysis which aims to minimize average polling delay while increasing network throughput. Based on the QoS requirements of each scheduling class, the priorities can be given between scheduling classes. However, this obtains reward only from high-class services because the priority does not differentiate the priorities of nodes.

Contention resolution has been discussed not only in IEEE 802.16 but also in IEEE 802.11. A classic Markov Chain model to analyze contention resolution in IEEE 802.11 has been proposed in (7). Because the bandwidth reservation is employed in the IEEE 802.16 standard, it is not practical for the SS in the IEEE 802.16 network to sense the medium status. Instead, the SS in the IEEE 802.16 network waits a fixed number of subsequent UL maps for receiving the response from the BS before entering into the back-off procedure. By considering this difference, a Markov model of contention in the IEEE 802.16 network is proposed in (6). This model consists of two types of states: back-off states and waiting states. The former illustrate the contention procedure. The latter represent the status that the SS waits for the response from the BS before entering into the back-off procedure. The parameters that control the contention resolution in the IEEE 802.16 network such as back-off start/end values have been investigated in (2). Moreover, the connections belonging to three types of scheduling classes (i.e. ertPS, nrtPS and BE) are able to join the contention resolution. The connection in each scheduling class has its own QoS requirements. However, there are no priorities employed in the contention resolution since the BS fixes the initial and maximum back-off window and each SS in the system uses the same value for all connections. In order to distinguish the priorities between the connections in different scheduling classes, a modified contention resolution process is proposed (8) to improve the system performance including end-to-end delay and throughput by assigning different initial window size to the connection in different scheduling class.

The research works summarized above provide the investigation of either unicast polling or contention resolution. However, the connections in the scheduling classes of ertPS, nrtPS and BE are allowed to use both bandwidth request mechanisms (i.e. unicast polling and contention resolution). Unfortunately, none of the research works shown above take this option into considerations. Their research is based on the assumption that only one bandwidth request mechanism is available. A research work considering both BR mechanisms is presented in (10). The authors first compare two bandwidth request mechanisms specified in the standard. Their results demonstrate that the contention resolution outperforms unicast polling when the probability of making bandwidth requests is low. However, the authors do not provide detailed analysis for each type of bandwidth request mechanisms. Moreover, the scheduling algorithms to help the BS make scheduling decisions are desired.

In this paper, two major contributions are included. First, a comprehensive study of both BR mechanisms is provided. We perform the performance analysis of each bandwidth request mechanism in terms of bandwidth utilization and delay. Second, two performance objectives are proposed. In order to achieve each of our proposed performance objectives, two scheduling algorithms are proposed to reach them individually. The simulation results presented in Section 2.6 show that our scheduling algorithms can also have the better performance while the corresponding performance objectives are satisfied.

2.4 Analytical Modeling

In this section, we analyze the performance of each BR mechanism in terms of the bandwidth utilization and the delay of delivering a BR. The network model used for analyzing both BR mechanisms is composed by a BS residing at the center of geographical area and N SSs randomly distributed in the service coverage of the BS. Each SS serves one identical variable bit rate (VBR) traffic, based on the traffic model introduced in (20), which is classified as a BE connection with the average probability Pr to transmit bandwidth requests. Additionally, we assume that each SS transmits at most one BR during the ex-

pected delay. This assumption is reasonable since there is no maximum delay requirement in BE connections and our objective is to make that the average delay is no more than the expected delay. Although piggybacking defined in the IEEE 802.16 standard is another way for SSs to transmit BRs, however, it is optional and not able to carry all types of BRs. Consequently, we do not consider piggybacking in this paper.

2.4.1 Unicast polling

We begin our analysis of unicast polling by investigating the minimum average number of unicast polling TxOPs allocated in each frame and the average delay of transmitting a BR. For ease of reference, a list of important notations are summarized in Table 2.1.

Notation	Description
N	Total number of SSs
N_p	Number of SSs with unicast polling TxOPs
FPS	Number of frames per second
M_p	The minimum average number of unicast polling TxOPs per frame
T_p	The expected delay
Pr	The probability of each SS to send BR
U_p	Bandwidth Utilization of unicast polling

Table 2.1 List of notations for unicast polling

Assume N_p is the total number of SSs assigned with unicast polling TxOPs, where $0 \leq N_p \leq N$. Since it is not necessary to schedule an unicast polling TxOP to each SS in every frame, we focus on the minimum number of unicast polling TxOPs which should be scheduled per frame in order to achieve the expected delay. Assume that the probability of the SS to make a BR is uniformly distributed between two consecutive unicast polling TxOPs. In order to maintain the expected delay, denoted as T_p, the BS has to schedule at least one unicast polling TxOP to the SS in every $2T_p$. Consequently, the minimum average number of unicast polling TxOPs assigned to each frame can be expressed as:

$$M_p \geq \frac{N_p}{2\mathtt{FPS}T_p} \tag{2.1}$$

where M_p stands for the minimum average number of unicast polling TxOP scheduled in each frame. Because of the nature of unicast polling, the unicast polling TxOP is wasted if the assigned SS does not transmitted a BR. Therefore, the bandwidth utilization of unicast polling is same as the probability of a SS to transmitted a BR (i.e. $U_p = Pr$).

2.4.2 Contention Resolution

Notation	Description
N	Total number of SSs
N_c	Number of SSs with contention TxOPs
FPS	The number of frames per second
M_c	The minimum average number of TxOPs for contention resolution per frame
T_c	The target delay of contention resolution
Pr	The probability of a SS to send BR
S	Back-off start value
E	Back-off end value
p	Probability of a unsuccessful transmission
W_S	Initial back-off window size
W_E	Maximum back-off window size
R	Maximum number of retries
q	Probability of the BS to accept a BR
$b(i, r_i)$	A back-off state in i-th attempt with random back-off counter r_i
$^{\dagger}w_1(i, t_i)$ $^{\ddagger}w_2(i, t_i)$	A waiting state in the branch of †collision/ ‡non-collision in i-th attempt and the SS has waited for t_i frames
p_f	the probability of failures
τ	The probability of a SS to transmit a BR in a randomly chosen TxOP
$^{\dagger}T_1$ $^{\ddagger}T_2$	The expected delay †before/‡after the contention window size reaches the W_E.
T_w	The maximum number of subsequent UL-MAP messages that a SS waits for a response from the BS

Table 2.2 List of notations for contention resolution

We analyze the contention resolution in IEEE 802.16 network by using a 2-D Markov

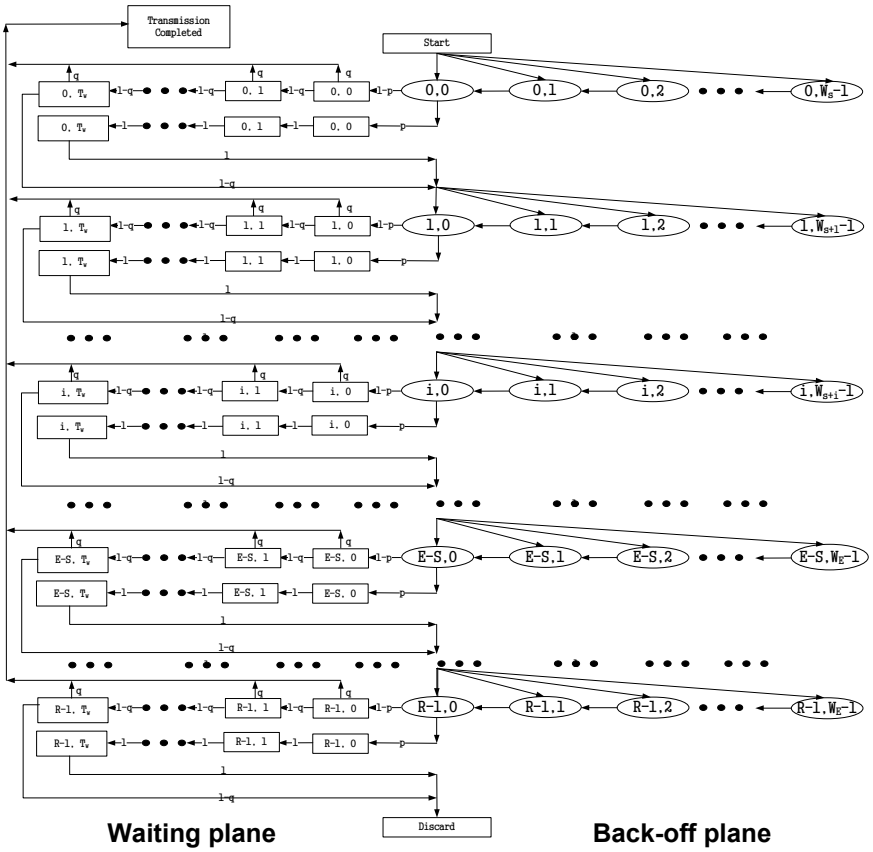


Figure 2.2 Markov Chain model for contention resolution

Chain (MC) model in Fig 2.2. As shown in the figure, each SS attempts to transmit a BR until the number of attempts reaches the maximum retry limit R. If the SS cannot transmit a BR successfully in R attempts, this BR shall be discarded. A list of important notations are summarized in Table 2.2.

According to the specification of contention resolution procedure described in the IEEE 802.16 standard, the SS shall select a random value within its back-off window. This random number indicates the number of contention TxOPs that the SS shall defer before transmitting a BR. After the contention transmission, the SS has to wait for a fixed number of subsequent UL maps before entering into the back-off procedure. Therefore, the contention resolution procedure is classified into two planes: back-off plane and waiting plane. The back-off plane describes how the SS transmits a BR (i.e. BEB in this paper). After transmitting a BR, the SS should wait for the response from the BS. The waiting plane is used to represent this waiting period. In Fig. 2.2 all states in back-off plane and waiting plane are denoted as ellipses and rectangles, respectively.

In back-off plane, each back-off state, denoted as $b(i, r_i)$, represents the i-th attempt of sending a BR with a random-chosen back-off counter r_i. This 2-D MC modeling is possible if we assume an independent and constant probability of an unsuccessful request, p, for each attempt. It is intuitive that this assumption results more accurate as long as the back-off window size, W, and the number of SSs with contention resolution TxOPs, N_c, get larger. The correctness of this assumption has been proven in (6). We refer to p as the *conditional collision probability* (7). A SS starts to transmit a BR when its back-off counter equals to 0 , regardless of the back-off stage. Once the independence is assumed, p is supported to be a constant value.

After a BR is transmitted, the SS enters into waiting plane which represents that the SS waits for a response from the BS. According to the IEEE 802.16 standard, the SS should consider that the transmission was failed if it does not receive a response from the BS within the number of subsequent UL-MAP messages specified by the parameter

of Contention-based Reservation Timeout. Here, we use T_w to represent the maximum number of subsequent UL-MAP messages that the SS can wait before entering into the back-off procedure. There are two possibilities that the SS cannot receive a response within T_w subsequent UL-MAPs: 1) the BR is collided with another BR sent from other SSs. 2) the BR is rejected by the BS. Based on these two possibilities, the waiting states are classified into two branches: collision and non-collision. The states in collision branch and non-collision branch are represented by $w_1(i, t_i)$ and $w_2(i, t_i)$, respectively, where i is the i-th attempt and t_i is the number of subsequent UL-MAP messages that the SS has waited after transmitting a BR.

As mentioned, p is the probability of an unsuccessful request. Thus, the probability to enter the branch of collision is also p. It can obtain that the probability of transition between all states in the branch of collision is 1 due to the failure of the BR transmission. Intuitively, the probability to enter the states in the branch of non-collision is $1 - p$. It is possible that the BS receives a BR successfully but rejects it due to the lack of radio resource or violating its scheduling policies. Suppose q is the probability of the BS to *accept* a BR in each frame. It is reasonable to assume that q is a constant for the waiting states of this 2-D Markov chain model. In fact, q is controlled by the policy of admission control and is independent of the operation of the MAC layer.

By combining these two factors which may cause failures of BR transmissions (collision and rejection by the BS), the *probability of failures*, denoted as p_f, can be represented as:

$$p_f = p + (1 - p)(1 - q)^{T_w} \tag{2.2}$$

Here, p is the probability of entering into the branch of collision. $(1 - p)(1 - q)^{T_w}$ denotes the probability of entering into the branch of non-collision but no response received from the BS. It leads to the following observation:

$$b(i, 0) = p_f \cdot b(i - 1, 0) \quad 0 < i \leq R \tag{2.3}$$

$$
\begin{cases}
P\{b(i,k) \mid b(i,k+1)\} = 1 \\
\qquad k \in (0, W_i - 2) \;\; i \in (0, m) & \text{(2.4a)} \\
P\{b(i+1,k) \mid b(i,0)\} = \dfrac{p_f}{W_{i+1}} \\
\qquad k \in (0, W_{i+1} - 1) \;\; i \in (0, E - S - 1) & \text{(2.4b)} \\
P\{b(i+1,k) \mid b(i,0)\} = \dfrac{p_f}{W_E} \\
\qquad k \in (0, W_E - 1) \;\; i \in (E - S, R - 1) & \text{(2.4c)} \\
P\{w_1(i,0) \mid b(i,0)\} = p \qquad i \in (0, R) & \text{(2.4d)} \\
P\{w_2(i,0) \mid b(i,0)\} = 1 - p \qquad i \in (0, R) & \text{(2.4e)} \\
P\{w_1(i,t_i+1) \mid w_1(i,t_i)\} = 1 \\
\qquad i \in (0, R), t_i \in (0, T_w - 1) & \text{(2.4f)} \\
P\{w_2(i,t_i+1) \mid w_2(i,t_i)\} = 1 - q \\
\qquad i \in (0, R), t_i \in (0, T_w - 1) & \text{(2.4g)} \\
P\{b(i+1,r_i) \mid w_1(i,T_w)\} = 1 \\
\qquad i \in (0, E - S - 1), r_i \in (0, W_i - 1) & \text{(2.4h)} \\
P\{b(i+1,r_i) \mid w_1(i,T_w)\} = 1 \\
\qquad i \in (E - S, R - 1), r_i \in (0, W_E - 1) & \text{(2.4i)} \\
P\{b(i+1,r_i) \mid w_2(i,T_w)\} = 1 - q \\
\qquad i \in (0, E - S - 1), r_i \in (0, W_i - 1) & \text{(2.4j)} \\
P\{b(i+1,r_i) \mid w_2(i,T_w)\} = 1 - q \\
\qquad i \in (E - S, R - 1), r_i \in (0, W_E - 1) & \text{(2.4k)}
\end{cases}
$$

Based on equation (2.2) and (2.3), the probabilities of transition between states shown in Fig 2.2 are summarized in equation (2.4a)−(2.4k). Equation (2.4a) represents the countdown of back-off counter. Equation (2.4b) and (2.4c) illustrate the probability of entering to each back-off state while the window size has and has not reached the maximum window size,

respectively. The probabilities of entering into the branch of collision and non-collision are shown in equation (2.4d) and (2.4e), respectively. Equation (2.4f) and (2.4g) are the probability between states between the branch of collision and non-collision, respectively. Equation (2.4h) and (2.4i) express that the SS enters into the back-off procedure from the branch of collision with different contention window size. Similarly, equation (2.4j) and (2.4k) express that the SS enters into the back-off procedure from the branch of non-collision with different contention window size.

Based on the size of contention window, the back-off states can be classified into two types; **Type 1**: the size of contention window is smaller than W_E. **Type 2**: the size of contention window has reached W_E. Suppose $b(i,k_1)$ and $b(j,k_2)$ denote the back-off states in Type 1 and Type 2, respectively. Additionally, $w_1(i,t_i)$ and $w_2(i,t_i)$ stand for the waiting states in the branch of collision and non-collision, respectively. Suppose P_{dis} represents the probability that a SS discards a BR because this BR cannot be transmitted successfully in R attempts. Thus, The sum of probability in all the states plus P_{dis} must equal to 1 as shown in equation (2.5). By simplifying equation (2.5), we can derive $b(0,0)$ shown in equation (2.6).

$$
\begin{aligned}
1 &= \sum_{i=0}^{E-S}\sum_{k_1=0}^{W_{S+i}-1} b(i,k_1) + \sum_{j=E-S+1}^{R-1}\sum_{k_2=0}^{W_E-1} b(j,k_2) \\
&+ \sum_{i=0}^{R-1}\sum_{t_i=0}^{T_w} w_1(i,t_i) \\
&+ \sum_{i=0}^{R-1}\sum_{t_i=0}^{T_w} w_2(i,t_i) + p_f \cdot b(R-1,0) \\
&= \sum_{i=0}^{E-S}\sum_{k_1=0}^{W_{S+i}-1} \left(\frac{W_{S+i}-k_1}{W_{S+i}} \cdot b(i,0)\right) \\
&+ \sum_{j=E-S+1}^{R-1}\sum_{k_2=0}^{W_E-1} \left(\frac{W_E-k_2}{W_E} \cdot b(j,0)\right) \\
&+ \sum_{i=0}^{R-1}\sum_{t_1=0}^{T_w} p \cdot b(i,0)
\end{aligned}
$$

$$
\begin{aligned}
&+ \sum_{i=0}^{R-1}\sum_{t_i=0}^{T_w}(1-p)(1-(1-q)^{t_i})b(i,0) \\
&+ p_f \cdot b(R-1,0) \\
&= \frac{b(0,0)}{2}\Bigg\{ \sum_{i=0}^{E-S} p_f^i(1+W_{s+i}) \\
&+ \sum_{j=E-S+1}^{R-1} p_f^j(1+W_E) + \sum_{i=0}^{R-1} 2T_w \cdot p \cdot p_f^i \\
&+ \sum_{i=0}^{R-1} \frac{2(1-p)(1-(1-q)^{T_w})}{q} p_f^i + 2P_f^{R+1} \Bigg\} \\
&= \frac{b(0,0)}{2}\Bigg\{ \Bigg[1 + 2p \cdot T_w \\
&+ \frac{2(1-p)(1-(1-q)^{T_w})}{q} \Bigg] \\
&\cdot \frac{1-p_f^R}{1-p_f} + \frac{1-(2p_f)^{E-S+1}}{1-2p_f} \cdot W_S \\
&+ 2^{E-S}W_S \frac{p_f^{E-S+1}-p_f^R}{1-p_f} + 2p_f^{R+1} \Bigg\}
\end{aligned}
\tag{2.5}
$$

$$
\begin{aligned}
b(0,0) &= 2\Bigg\{ \Bigg[1 + 2p \cdot T_w + \frac{2(1-p)(1-(1-q)^{T_w})}{q} \Bigg] \\
&\cdot \frac{1-p_f^R}{1-p_f} + \frac{1-(2p_f)^{E-S+1}}{1-2p_f} \cdot W_S \\
&+ 2^{E-S}W_S \frac{p_f^{E-S+1}-p_f^R}{1-p_f} + 2p_f^{R+1} \Bigg\}^{-1}
\end{aligned}
\tag{2.6}
$$

The probability that a SS transmits a BR in a randomly chosen contention TxOP can be calculated as the sum of $b(i,0)$, where $0 \leq i \leq R-1$. This probability, denoted as τ, is expressed as:

$$
\begin{aligned}
\tau &= \sum_{i=0}^{R-1} b(i,0) \\
&= \frac{b(0,0)}{1-p_f}
\end{aligned}
\tag{2.7}
$$

As shown in equation (2.3), $b(0,0)$ is represented as a function of P_f which is a function of p presented in equation (2.2). Thius, the value of τ stated in equation (2.7) can be

expressed as a function of the conditional collision probability, p, which is unknown in our model. Again, p is the probability that a collision occurs, which is equivalent to the probability of at least two SSs transmitting BRs at the same contention TxOP. Thus, p can be represented as

$$p = 1 - (1 - \tau)^{(N_c \cdot Pr)-1} \tag{2.8}$$

where τ is the probability that a SS transmits a BR at the randomly chosen contention TxOP shown in equation (2.7).

By using equation (2.7) and (2.8), we can solve these two unknown values, p and τ, based on the known value of back-off start and end value (i.e. S and E), the probability of a SS to send a BR (i.e. Pr), the probability of a BS to accept a BR (i.e. q) and the number of SSs with contention TxOPs (i.e. N_c).

To analyze the bandwidth usage of contention TxOPs, it is necessary to find the bandwidth utilization, U_c, which is defined as the ratio of the number of TxOPs which deliver BRs successfully to the total number of contention TxOPs. To get this ratio, first we investigate the *probability of transmission*, denoted as p_{tx}, which is referred to the probability that at least one SS transmitting a BR at a TxOP. This probability can be obtained as:

$$p_{tx} = 1 - (1 - \tau)^{N_c \cdot Pr} \tag{2.9}$$

The probability of a successful transmission, denoted as p_{st}, is the probability that a BR is delivered successfully and the BS grants this BR. This probability can be achieved by using conditional probability that only one SS transmits a BR at a TxOP *and* the BS has enough bandwidth to serve this BR under the condition that at least one transmission is transmitted at this TxOP. Therefore, the probability of a successful transmission can be addressed as:

$$p_{st} = \frac{n\tau(1 - \tau)^{(N_c P_r)-1}}{p_{tx}}(1 - (1 - q)^{T_w}) \tag{2.10}$$

From equation (2.9) and (2.10), the probability of a TxOP which delivers a BR success-

fully, represented as p_{sbr}, is derived as:

$$p_{sbr} = p_{st} \cdot p_{tx} = n\tau(1-\tau)^{(N_c Pr)-1}(1-(1-q)^{T_w}) \tag{2.11}$$

Intuitively, the probability that a BR is delivered in a given TxOP is equivalent to the probability of a TxOP to be utilized successfully. Consequently, the bandwidth utilization of contention TxOPs, U_c is same as p_{sbr}.

Although the maximum delay requirement is not a necessary requirement for BE connections, in practice, we still hope the delay can be limited into certain bound which is considered as our expected delay, T_c. Here, the delay is calculated as the time difference between the time that the SS intends to send a BR and the time that the SS receives a response from the BS. One of the important factors to affect the delay is the number of contention TxOPs scheduled by the BS in each frame. In this paper, we focus on the relation between the minimum average number of contention TxOPs assigned per frame (denoted as M_c) and the target delay (denoted as T_c). Based on the contention windows size, the expected delay can be calculated into two sections: 1) $i \leq E - S$ and 2) $E - S < i \leq R$, where i is the i-th attempt. Let T_1 stand for the expected delay in the first section. It can be calculated as equation (2.12). Similarly, the delay of the second section, denoted as T_2, can be derived as equation (2.13). It is intuitive that the sum of the delay of two sections is at most the target delay which is represented as T_c. Moreover, in equation (2.12) and (2.13), everything is known except M_c and p which are the minimum average number of contention TxOPs assigned per frame and the probability of an unsuccessful transmission, respectively. Therefore, we can use $H(M_c, p)$ to represent the total delay as the sum of delay in these two sections. By writing formally, it can be expressed as equation (2.14).

$$\begin{aligned} T_1 &= \sum_{j=S}^{E} p_f^{j-S}(1-p_f)\left[\frac{1}{W_j \mathtt{FPS}} \cdot \sum_{k=0}^{W_j-1}\left\lceil \frac{k}{M_c} \right\rceil\right] \\ &+ \sum_{i=0}^{T_w-1} \frac{i}{\mathtt{FPS}} q(1-q)^i \end{aligned}$$

$$+ \sum_{m=S}^{j-1} \left(\frac{1}{W_m \texttt{FPS}} \sum_{k=0}^{W_m-1} \left\lceil \frac{k}{M_c} \right\rceil + \frac{T_w}{\texttt{FPS}} \right) \Bigg] \tag{2.12}$$

$$\begin{aligned} T_2 = & \sum_{n=E-S+2}^{R} p_f^{n-1}(1-P_f) \Bigg[\frac{1}{W_E \texttt{FPS}} \sum_{k=0}^{W_E-1} \left\lceil \frac{k}{M_c} \right\rceil \\ & + \sum_{i=0}^{T_w-1} \frac{i}{\texttt{FPS}} q(1-q)^i \\ & + \sum_{d=E_S+2}^{n-1} \left(\frac{1}{W_E \texttt{FPS}} \sum_{k=0}^{W_E-1} \left\lceil \frac{k}{M_c} \right\rceil + \frac{T_w}{\texttt{FPS}} \right) \\ & + T_b \Bigg] \end{aligned}$$

where

$$T_b = \sum_{m=S}^{E-S+1} \left(\frac{1}{W_m \texttt{FPS}} \sum_{k=0}^{W_m-1} \left\lceil \frac{k}{M_c} \right\rceil + \frac{T_w}{\texttt{FPS}} \right) \tag{2.13}$$

$$T_c \geq T_1 + T_2 = H(M_c, p) \tag{2.14}$$

2.5 Scheduling Algorithms for performance objectives

Based on the analysis shown in section 2.4, we proposed two scheduling algorithms to meet the two performance objectives proposed in this paper, respectively:

1. Maximize the bandwidth utilization under the condition of satisfying a given target delay requirement (represented as *Fixed-delay-MAX-Utilization* in the rest of this section)

2. Minimize the target delay when a given bandwidth utilization as a constraint is given (represented as *Fixed-Utilization-MIN-delay* in the rest of this section)

To meet the first objective, a scheduling algorithm, called *Maximum Bandwidth Utilization Scheduling Algorithm* (MAX-U), is proposed. It helps the BS to schedule the number of TxOPs and the number of participating SSs for each BR mechanism in order to maximize the bandwidth utilization while satisfying the target delay. Similarly, the scheduling algorithm proposed for the second objective is called *Minimize Delay Scheduling Algorithm* (MIN-D). It helps the BS to find the combination of TxOPs assigned for each BR mechanism such that the system delay is minimized while maintaining the desired utilization. Note that both scheduling algorithms help the BS schedule either unicast polling TxOPs or contention TxOPs to each SS to achieve the corresponding performance objective. No SSs receive both types of TxOPs at the same time.

In this paper, we only focus on the BR mechanisms. Thus, the bandwidth utilization indicated in this paper is the bandwidth utilization of TxOPs assigned for both BR mechanisms (i.e. unicast polling and contention resolution). Moreover, the TxOPs scheduled for each mechanism are only used for transmitting BR messages.

2.5.1 MAX-U

This algorithm is designed to satisfy our first performance objective: maximize the bandwidth utilization while satisfying the fixed delay requirement. The flow of this algorithm is shown in Fig. 2.3. Suppose T_D is the given achievable target delay. Our objective is to find the number of unicast polling TxOPs and contention TxOPs scheduled in each frame such that the bandwidth utilization is maximized.

In this algorithm, each SS is scheduled with either one of the BR mechanisms: unicast polling TxOPs and contention TxOPs. Suppose N_c and N_p are the number of SSs scheduled with contention and unicast polling TxOPs, respectively, such that $N_c + N_p = N$. The objective of the algorithm is to maximize the bandwidth utilization while achieve the given target delay. For all combinations of (N_p, N_c), we calculate the corresponding value of M_p and M_c for each combination and select a combination of (N_p, N_c) and the corresponding M_p

Algorithm 1 MAX-U

Input: All variables specified in Table 2.1 and 2.2

Output: N_p is the number of SSs with unicast polling TxOPs.
N_c is the number of SSs with contention resolution TxOP.
M_p is the average number of TxOPs scheduled for unicast polling per frame.
M_c is the average number of TxOPs scheduled for contention resolution per frame.

For i =0 to N **do**
$N_p^i \leftarrow i, \quad N_c^i \leftarrow N - N_p$
Unicast Polling:
$M_p^i \leftarrow M_p^i$ calculate by equation (2.1)
Contention Resolution:
a. Solve τ^i and p^i by using equation (2.7) and (2.8) with given N_c^i.
b. $M_c^i \leftarrow M_c^i$ calculated by equation (2.14) with a known p^i.
c. $p_{sbr}^i \leftarrow p_{sbr}^i$ calculated by equation (2.11).
Finalize:
$U_t^i \leftarrow \frac{M_p^i Pr + M_c^i p_{sbr}^i}{M_p^i + M_c^i}$
End For
$U_t \leftarrow Max\{U_t^i\}$ with $Min\{M_c^i + M_p^i\}$
$M_p \leftarrow M_p^i$, $N_p \leftarrow N_p^i$, $M_c \leftarrow M_c^i$, $N_c \leftarrow N_c^i$
Return M_p N_p M_c N_c

Step 1	Find all combinations of (N_p, N_c) such that $N_p + N_c = N$.
Step 2	For each (N_p, N_c), calculate the corresponding (M_p, M_c) and the bandwidth utilization while the target delay, T_D, is satisfied.
Step 3	Return the (N_p, N_c) and the corresponding (M_p, M_c) such that the bandwidth utilization is maximized.

Figure 2.3 The steps of MAX-U

and M_c, which can maximize the bandwidth utilization. Note that the overall throughput may be higher if we can minimize the number of TxOPs assigned for BR mechanisms because there is more bandwidth which can be assigned for data transmissions. Therefore, if there are multiple combinations which result the same maximum bandwidth utilization, the one with minimum number of TxOPs (i.e. M_c+M_p) is selected.

Algorithm 2 MIN-D

Input: All variables specified in Table 2.1 and 2.2

Output: N_p is the number of SSs with unicast polling TxOPs.
N_c is the number of SSs with contention resolution TxOPs
M_p is the average number of unicast polling TxOPs per frame.
M_c is the average number of contention resolution TxOPs per frame.

For $i = 0$ to N **do**
$N_p^i \leftarrow i \quad N_c^i \leftarrow N - N_p$
$K^i \leftarrow$ the set of all combinations of (M_p^i, M_c^i) such that equation (2.15) is satisfied and $M_p^i \leq N_p^i$.
For j = 1 to $|K^i|$
Unicast Polling:
a. $T_p^j \leftarrow$ Calculated by equation (2.1).
Contention Resolution:
a. Solve τ^i and p^i by using equation (2.7) and (2.8) with given N_c^i.
b. $T_c^j \leftarrow T_c^j$ Calculated by equation 2.14.
End For
$T_D^i \leftarrow Min\{Max\{T_p^j, T_c^j\}\}$
$M_p^i \leftarrow M_p^j$
$M_c^i \leftarrow M_c^j$
End For
$T_D \leftarrow Min\{T_D^i\}$, $M_p \leftarrow M_p^i$, $N_p \leftarrow N_p^i$, $M_c \leftarrow M_c^i$, $N_c \leftarrow N_c^i$
Return $M_p \;\; N_p \;\; M_c \;\; N_c$

Step 1	Find all combination of (N_p, N_c) such that $N_p + N_c = N$.
Step 2	$\forall$ (N_p, N_c), find all corresponding (M_p, M_c) such that $M_p(1 - P_r) + M_c(1 - p_{sbr}) \leq S_u$ and $M_p \leq N_p$.
Step 3	$\forall$ (N_p, N_c), find the corresponding delay of each (M_p, M_c) and select one with minimum delay.
Step 4	$\forall$ (N_p, N_c), set the delay as the corresponding delay of the picked (M_p, M_c).
Step 5	Return the (N_p, N_c) with minimum delay.

Figure 2.4 The steps of MIN-D

2.5.2 MIN-D

This algorithm focus on achieving our second performance objective: minimizing the delay while satisfying a given bandwidth utilization requirement. The detail steps of this algorithm are presented in Fig. 2.4. Assume U_t is the given bandwidth utilization with the fixed number of TxOPs, S_t, for both BR mechanisms (i.e. $M_p + M_c = S_t$). Thus, the number of unused TxOPs, denoted as S_u, can be represented as:

$$S_u = (1 - U_t)S_t$$

It is intuitive that the total unused TxOPs of both BR mechanisms are at most S_u. Formally, it can be expressed as:

$$M_p(1 - Pr) + M_c(1 - p_{sbr}) \leq S_u \tag{2.15}$$

Similar to Algorithm 1, we exam all combinations of (N_p, N_c) such that $N_p + N_c = N$. Our objective is to find a combination of (N_p, N_c) with the minimum overall expected delay while equation (2.15) is satisfied.

For each pair of (N_p, N_c), there exist several pairs of (M_p, M_c) which satisfy the constraint stated in equation (2.15). Suppose M' is the set of qualified (M_p, M_c) for each pair of (N_p, N_c). Therefore, we check all combinations of $(M_p, M_c) \in M'$ and find the combinations resulting the delay is minimized as our candidates. Here, delay is defined as $\max\{T_p, T_c\}$, where T_p and T_c are the delay caused by unicast polling and contention resolution, respec-

tively. Consequently, for each pair of (N_p, N_c), there are at least one pair of (M_p, M_c) as our candidates. Among these candidates, we pick one candidate with the minimum delay as our solution for the scheduling decision.

2.6 Numerical and Simulation Results

2.6.1 System Set Up

In this section, we validate the theoretical results with our simulation results. The theoretical results are made by the *Matlab 2009a*. The simulation results are conducted by our simulator. The simulator is written in C and followed the IEEE 802.16 standard closely. Both analytical and simulation results are also compared with two ordinary schemes: *1)* Unicast Polling only. *2)* Contention Resolution only. Table 2.3 summarizes the system parameters used in our numerical analysis and simulation. In our simulation, each SS serves one HTTP web browsing traffic (12) (13) which is classified as a BE connection. In order to increase the variety of BE traffics, the mean packet size is randomly selected from 512 to 1024 bytes. Because the mean traffic rate is fixed, the mean traffic rate can be calculated based on the selected mean packet size.

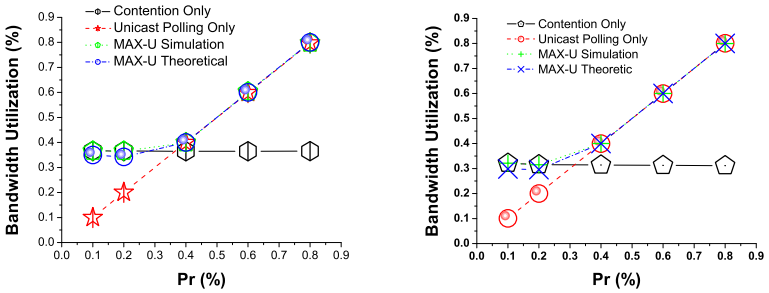
Parameters	**Value**
Number of BS	1
Number of SS	200, 300, 400, 500
Frame Duration	20 ms
Modulation	BPSK, QPSK, 16QAM, 64QAM
TTG/RTG	10 μs
SSTG	4 μs
Application	HTTP
Traffic Type	VBR
Scheduling Class	BE
Mean Packet Size	512 $\sim$ 1024(byte)
Mean Traffic Rate	2Kbps

Table 2.3 Simulation Parameters

2.6.2 MAX-U

The target delay used in this simulation is 1 second which is the most common delay used for BE connections. The results of bandwidth utilization under different Pr are shown in Fig. 2.5(a) and Fig. 2.5(b). It is easy to observe that the results of bandwidth utilization are similar with different number of SSs. It shows that the bandwidth utilization does not strongly relate to the number of SSs in the system. The utilization of contention resolution only is always has around 35 % no matter what value of Pr is. On the other hand, the utilization of unicast polling only is very close to the value of Pr. By these results, we can conclude that the unicast polling can achieve better bandwidth utilization if Pr is larger than 0.35. Therefore, it is impossible to always reach the better performance if only one BR mechanism is considered.

As shown in the figures, our analytic and simulation results are very close to each other. This validates this analysis presented in Section 2.4. Additionally, our results always achieve the better bandwidth utilization produced by either unicast polling only or contention resolution only. For example, in Fig. 2.5(a), our algorithm achieves around 35 % of the bandwidth utilization when $Pr = 0.1$, which is similar to the one that contention only achieves. However, unicast polling only results 10 % of bandwidth utilization. When $Pr = 0.8$, both unicast polling and our algorithm reach 80 % of bandwidth utilization. The contention only still keeps its bandwidth utilization around 35 %. It is because our scheduling algorithm (i.e. MAX-U) can help the BS schedule one type of BR mechanisms which can achieve better performance according to the current network status. It is worth to note that our scheduling algorithm (MAX-U) schedules all SSs with either unicast polling TxOPs or contention TxOPs. The combinations in between (i.e. part of SS with unicast polling TxOPs and the rest of them with contention TxOPs) do not exist. It is because the contention resolution can always give 35 % of bandwidth utilization and it will be chosen if the unicast polling cannot contribute as high bandwidth utilization as it does. On the other hand, the unicast polling will always be chosen when it can have more than 35 % in



(a) Number of SS = 200 (b) Number of SS = 500

Figure 2.5 Simulation Results of *MAX-U*

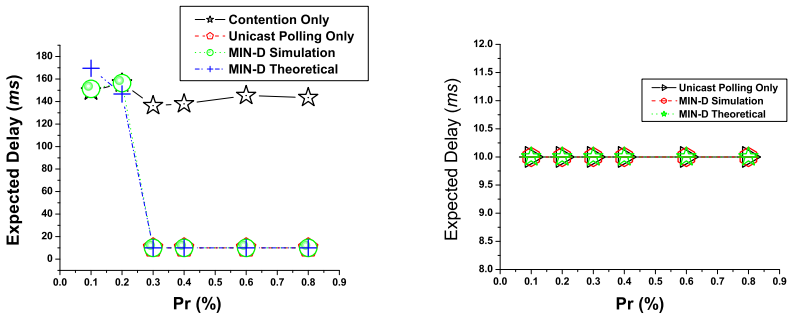
bandwidth utilization (i.e. $Pr > 35$ %).

2.6.3 MIN-D

Fig. 2.6(a) and Fig. 2.6(b) show the relationship between the expected delay and Pr while the target bandwidth utilization is 0.3 and 0.5, respectively. From the figures, we obtain that our scheduling algorithm (i.e. MIN-D) always picks a BR mechanism resulting better performance (i.e. shorter delay). For instance, in Fig. 2.6(a), both unicast polling and our algorithm reach 10 ms of delay when $Pr = 0.4$. However, the contention only keeps the delay around 145 ms in all values of Pr. In Fig. 2.6(a), there are no results for unicast polling only when $Pr = 0.1$ and 0.2. It is because the bandwidth utilization cannot achieve 0.3 if only unicast polling is used. Similarly, there are no results for contention only in Fig. 2.6(b) since the contention resolution cannot reach 50 % of bandwidth utilization.

Pr	N_p	M_p	N_c	M_c
0.1	2	1	498	15
0.2	0	0	500	15
0.3	500	500	0	0
0.4	500	500	0	0
0.6	500	500	0	0
0.8	500	500	0	0

Table 2.4 Simulation results of MIN-D



(a) U = 0.3 (b) U = 0.5

Figure 2.6 Simulation Results of *MIN-D*

Table 2.4 shows the simulation results of the scheduling algorithm in terms of the number of SSs and the number of TxOPs assigned to each BR mechanism. Here, the target bandwidth utilization is 0.3. It is worth to note that both BR mechanisms are scheduled for BR transmissions when $Pr = 0.1$. It is because the performance requirement (i.e. U = 0.3) cannot be achieve if only one BR mechanisms is considered. This result shows an example that the better performance can be achieved by scheduling both types of BR mechanisms.

2.7 Conclusion

According to the IEEE 802.16 standard, the connections belonging to ertPS, nrtPS and BE are allowed to make bandwidth requests (BRs) via both BR mechanisms (i.e. unicast polling and contention resolution). The mechanism that the BS schedules to those connections may result different system performance because of the nature of each BR mechanism. However, most conventional research works limit the option to consider only one type of BR mechanisms. A scheduling scheme by considering both types of BR mechanisms is desired for the BS in order to optimize the system performance. Besides, it is not necessary for BS to perform either unicast polling or contention resolution to all SSs within one frame. Instead, the BS needs to schedule the appropriate number of contention resolution or unicast polling TxOPs to the SS in order to meet the delay requirement. Therefore, the scheduling

decision should be made in a multi-frame basis.

In this paper, we provide the performance analysis of each BR mechanisms in terms of bandwidth utilization and expected delay. Based on the analysis, we take both BR mechanisms into account and propose two scheduling algorithms to help the BS make the scheduling decision based on the current network status such that the corresponding performance objectives are achieved. There are two performance objectives proposed in this paper: 1) Maximizing the bandwidth utilization under the condition that the target delay is satisfied. 2) Minimizing the delay while the desired bandwidth utilization is reached. Our numerical and simulation confirm that the scheduling algorithms can always have the better performance by scheduling the number of transmission opportunities to one of the BR mechanism. Additionally, when the probability of making BR (i.e. Pr) is 0.1, a hybrid decision (i.e. Scheduling SSs with two BR mechanisms) can conduct the minimum delay while satisfying the desired bandwidth utilization.

CHAPTER 3. Bandwidth Recycling in IEEE 802.16 Networks

A paper to be published in IEEE Transactions on Mobile Computing

Volume: 9 , Issue: 10 Page(s): 1451 - 1464

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 guaranteeing 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 guaranteeing 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% 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% when the network is in the steady state.

3.1 Introduction

The IEEE 802.16 standards (e.g., 802.16-2004 (1), 802.16e (53)) 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 guaranteeing 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 guaranteeing 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 guaranteeing 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 guaranteeing services, our research objective is twofold: 1) we do not change the existing bandwidth reservation to maintain the same QoS guaranteeing 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 guaranteeing 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 guaranteeing services.

According to the IEEE 802.16 standard, SSs scheduled on the uplink (UL) map should have transmission opportunities in the current frame. Those 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. 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 a RM successfully. Our theoretical analysis shows that the CS has at least 42% of probability to receive a 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. 2) the CS does not have non-real time data to transmit while receiving a RM. To mitigate those factors, additional scheduling algorithms are proposed. Our simulation results show that the proposed can further improve the average throughput by 40% when the network is in the steady state (i.e., 15∼75 second in our simulation).

The rest of this paper is organized as follows. In Section 3.2, we provide background information of IEEE 802.16. Motivation and related works are presented in Section 3.3. Our proposed scheme is presented in Section 3.4. The analysis of the proposed scheme and simulation results are placed in Section 3.5 and Section 3.6. In Section 3.7, 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 3.8. At the end, the conclusion is given in Section 3.9.

3.2 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 Orthogonal Frequency-Division Multiple Access (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 uplink (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 can not 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 a MAC frame.

The structure of a 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 a MAC frame.

The IEEE 802.16 network is connection-oriented. It gives the advantage of having better control over network resource to provide QoS guaranteeing 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 establish 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 be only 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.

3.3 Motivation and Related Work

Bandwidth reservation allows IEEE 802.16 networks to provide the QoS guaranteing 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 pre-assigned 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 (16), a dynamic resource reservation mechanism is proposed. It can dynamically 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 (15). 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 signalling cost. In (17), the authors enhanced the system throughput by using concur-

rent transmission in mesh mode. The authors in (18) 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 (19). 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 guaranteeing services and introducing no extra delay.

3.4 Proposed Scheme

The objectives of our research are twofold: 1) The same QoS guaranteeing services are provided by maintaining the existing bandwidth reservation. 2) the bandwidth utilization is improved by recycling the unused bandwidth. To achieve these objectives, our scheme named *Bandwidth Recycling* is proposed. The main idea of the proposed scheme is to allow the BS to pre-assign 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 pre-assigned CS and TS. As shown in Fig. 3.1, 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.

The UL map including burst profiles and offsets of each TS is received by all SSs within the network. Thus, if a 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

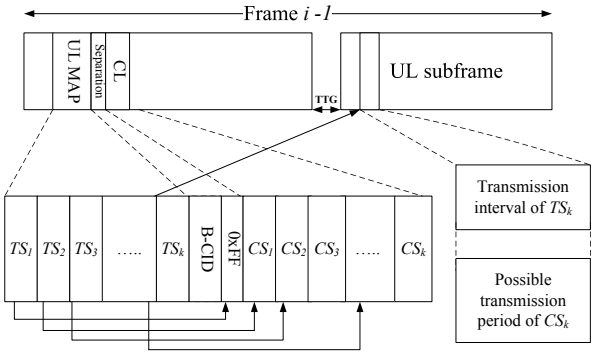


Figure 3.1 The mapping relation between CSs and TSs in a MAC frame

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. 3.1, 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 3.4.1. 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.

3.4.1 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 stuffed byte value (SBV). If the size of the unused region is at least the size of a 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.

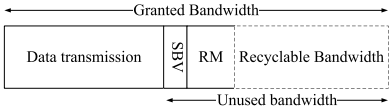


Figure 3.2 Messages to release the unused bandwidth within a UL transmission interval.

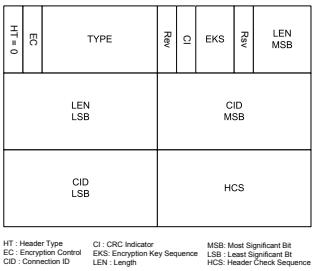


Figure 3.3 The format of RM

Instead of padding all portion of the unused bandwidth in our scheme, a TS with unused bandwidth transmits only a SBV and a releasing message (RM) shown in Fig. 3.2. 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 MAC PDU with no payload (6 bytes) shown in Fig. 3.3. 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.

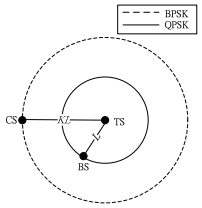


Figure 3.4 An example of corresponding locations of TS, BS and CS.

For example, Fig. 3.4 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 dished circle. The radius of the dished 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 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 a RM is received from its corresponding TS. Once a 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.

3.4.2 Scheduling Algorithm

Assume 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 can not be performed simultaneously, we can not 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 3 is used to schedule a 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 non-zero CG and BE with non-zero 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.

3.5 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 a 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

Algorithm 3 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

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 a RM successfully. Obviously, this probability affects the bandwidth recycling rate (BBR). BBR stands for the percentage of the unused bandwidth which is recycled. 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 3.6.

3.5.1 Analysis of Potential Unused Bandwidth

Based on the traffic generation rate, the applications can be classified into two types: constant bit rate (CBR) and variable bit rate (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 traffics. However, VBR applications generate data in a variable rate. It is hard for a 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 guaranteeing 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 (20), 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 guaranteeing 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}) (21) (22). 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:

$$W = BD_f \leq W_{max} \tag{3.1}$$

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 inter-frame 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 the following condition:

$$X_{i-1} < W_i - \max\{0, Q_{i-1} - W_{i-1}\} \tag{3.2}$$

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:

$$P_u(i) = \int_0^{X_{i-1}} p(X)dX \tag{3.3}$$

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

$$E(i) = \int_0^{X_i - 1} Xp(X)dX \tag{3.4}$$

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:

$$R_u = \frac{\sum_{i=0}^{N-1} E(i)}{\sum_{i=0}^{N-1} W_i} \tag{3.5}$$

3.5.2 The probability of RMs received by the corresponding CSs successfully

Assume 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

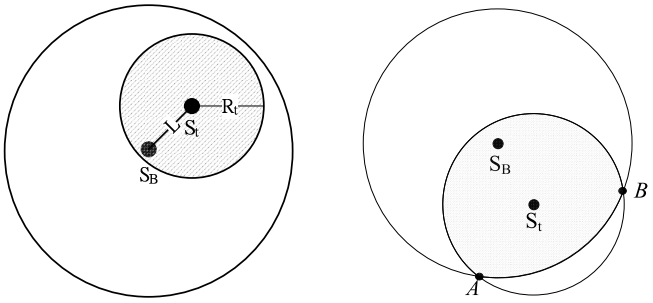
transmission rate of each SS may be variant depending on the PHY transmission technology and transmission power. For a given SS, S_t, let $R_t^{(B)}$, $R_t^{(Q)}$, $R_t^{(16)}$ and $R_t^{(64)}$ denote as the transmission range of BPSK, QPSK, 16-QAM and 64-QAM, respectively. In our scheme, the RM should 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:

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

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 a 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 a 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 a RM is equivalent to the transmission coverage of a RM overlapping with the service coverage of S_B. We analyze the average probability that the CS can receive a RM successfully.



(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

Figure 3.5 Possible geographical relationship between S_t and S_B

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. 3.5(a). 2) only part of the coverage of S_t is within the service coverage of S_B, shown as Fig. 3.5(b). The coverage of S_t means the maximal coverage of RMs transmitted by S_t. The analysis of each category is presented as follows.

3.5.2.1 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:

$$A_{in} = \pi R_t^2 \tag{3.6}$$

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:

$$P_c(t) = \frac{R_t^2}{R^2} \tag{3.7}$$

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:

$$L \leq R - R_t \tag{3.8}$$

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

$$R_t = KL \tag{3.9}$$

where K is a constant depending on the transmission power and modulation that S_t uses to communicate with the S_B. By combining equations (3.8) and (3.9), S_t belongs to this category if:

$$L \leq \frac{R}{K+1} \tag{3.10}$$

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

$$P_{oc}(t) = \frac{1}{(K+1)^2} \tag{3.11}$$

3.5.2.2 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. 3.5(b). Based on the location of S_t, we can classify into two cases:

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

Fig. 3.6 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, the calculation is omitted from this paper. The total area, A_{total}, can be presented as:

$$A_{total} = R^2\theta + R_t^2\alpha - LL_2 \tag{3.12}$$

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

$$P_s(t) = \frac{R^2\theta + R_t^2\alpha - LL_2}{\pi R^2} \tag{3.13}$$

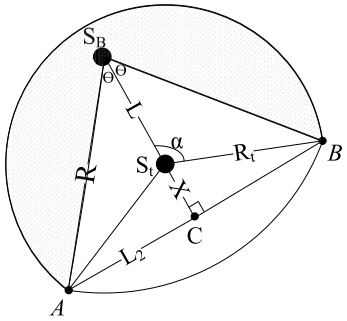


Figure 3.6 Both S_B and S_t are in the same side of $\overline{AB}$

In this case, the borders of both S_t and S_B coverage must intersect on two points. From equation (3.10), 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:

$$L \leq \frac{R}{\sqrt{1+K^2}} \tag{3.14}$$

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:

$$P_{os}(t) = \frac{2K}{(K+1)^2(1+K^2)} \tag{3.15}$$

II. S_B and S_t are on different side of $\overline{AB}$:

Fig. 3.7 illustrates the overlapping coverage of S_t and S_B. Each of them locates on one side of $\overline{AB}$. The total area, A'_{total}, that S_j can receive the RM is:

$$A'_{total} = R^2\beta + R_i^2\lambda - LL_4 \tag{3.16}$$

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

$$P_e(t) = \frac{R^2\beta + R_i^2\lambda - LL_4}{\pi R^2} \tag{3.17}$$

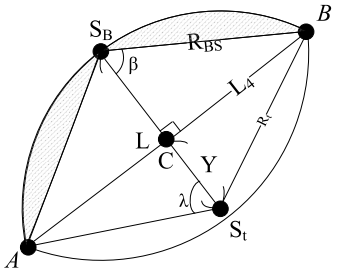


Figure 3.7 S_B and S_t are in each side of $\overline{AB}$.

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 equation (3.14), 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 no longer than R. By calculating the ring area between lower bound and upper bound of L, the probability of S_t belonging to this case, $P_{oe}(t)$, can be derived as:

$$P_{oe}(t) = \frac{K^2}{1+K^2} \tag{3.18}$$

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

$$P_t(t) = P_e(t)P_{oe}(t) + P_s(t)P_{os}(t) + P_c(t)P_{oc}(t) \tag{3.19}$$

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

$$P_t = \frac{\sum_{t=1}^{\|T\|} P_t(t)}{\|T\|} \tag{3.20}$$

where T is the set of all TSs.

3.5.3 Performance analysis of proposed scheme

Assume 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:

$$P_{CL}(n) = \begin{cases} \frac{\|Q_{CL}\|}{\|Q_n\|} & \|Q_n\| \geq \|Q_{CL}\| \\ 1 & \text{Otherwise} \end{cases} \tag{3.21}$$

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 inter-frame dependence, it can be expressed as the following condition:

$$Y_{i-1} > W_i^{nrt} - \max\{0, Q_{i-1}^{nrt} - W_{i-1}^{nrt}\} \tag{3.22}$$

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 :

$$P_u(u) = \int_{Y_{i-1}}^{\lambda_{max}^{nrt}} P(X)dX \tag{3.23}$$

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 a RM must be scheduled on the CS and have non-real time data to be transmitted. From equations (3.21) and (3.23), the probability that a CS satisfies these two conditions can be derived as:

$$P_r = \frac{\sum_{j=1}^{\|Q_n\|} P_u(j)(P_{CL}(j))}{\|Q_n\|} \tag{3.24}$$

If the CS recycles the unused bandwidth successfully, then it must meet the three conditions: 1) a 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 equations (3.20) and (3.24), the

recycling rate ,defined as the average probability that a CS recycles the unused bandwidth successfully, can be obtained as:

$$P_{recycle} = P_r P_t \tag{3.25}$$

Suppose B_g is the total bandwidth in the system and the unused bandwidth of the system is B_w. By equation (3.25), The total throughput gain, TG, can be derived as:

$$TG = \frac{P_{recycle} B_w}{B_g - B_w} \tag{3.26}$$

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.

3.5.4 Overhead analysis of 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. 3.1, the separation contains a broadcast CID (B-CID) and a SBV (0xFF). It costs 3 bytes of overhead (16 bits for B-CID and one byte for SBV). In addition, The CL is composed by 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:

$$OH_{DL} \leq 3 + 7B_{TS} \tag{3.27}$$

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. A RM is composed by 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 an UL subframe is calculated as:

$$OH_{UL} \leq 6B_{TS} \tag{3.28}$$

where B_{TS} is the number of TSs scheduled on the UL map. From equation (3.27) and (3.28), the total overhead introduced by our scheme in a MAC frame is concluded as:

$$OH = OH_{DL} + OH_{DL} \leq 3 + 7B_{TS} + 6B_{TS} \tag{3.29}$$

3.5.5 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 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 3.6.

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 equation

(3.23), is small. It leads to low P_r (from equation (3.24)). 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.,

$$B_{all} = B_{rt} + Bnrt$$

   In this stage, the 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 (equation (3.23)) may be higher than the stage 1 but still limited. Based on equation (3.24), (3.25) and (3.26), 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 Stage 1 and 2.

4. Stage 4 (full load) :

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

3.5.6 Tradeoff

In the IEEE 802.16 standard, the SS can adjust the amount of reserved bandwidth via BRs. In this subsection, 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 transmits. 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 a SS. The SS has to send m BRs which are $19m$ bytes (considering standard 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 a 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 tradeoff 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 a 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.

3.6 Simulation Results

Our simulation is conducted by using Qualnet 4.5 (23), 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.

3.6.1 Simulation Model

Our simulation model is composed by 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 3.1. 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.

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

Table 3.1 The system parameters used in our simulation

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 multi-media 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 (24) (25) (26). 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 3.2.

In our simulation, each SS serves at least one and up to 5 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 3.2 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 1024 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 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 0 to 15th second. Moreover, the real time connection stops to generate data from 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.

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				

Table 3.2 The traffic model used in the simulation

3.6.2 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:

   $$TG = \frac{T_{recycle} - T_{no\_recycle}}{T_{no\_recycle}}$$

   where $T_{recycle}$ and $T_{no\_recycle}$ 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 for-

mally as:

$$UBR = \frac{B_{unused\_bw}}{B_{total\_bw}}$$

where B_{unused_bw} and B_{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:

$$BRR = \frac{B_{recycled}}{B_{unused\_bw}}$$

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

3.6.3 Simulation Results

Fig. 3.8 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%. 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 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. 3.9. 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

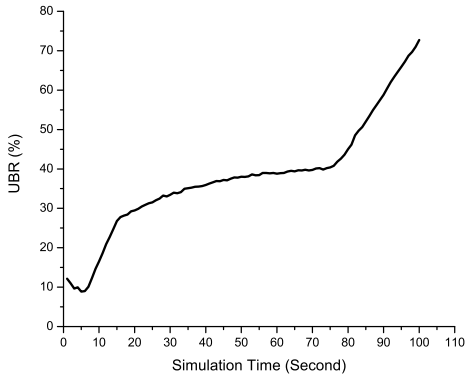


Figure 3.8 Simulation results of *UBR*

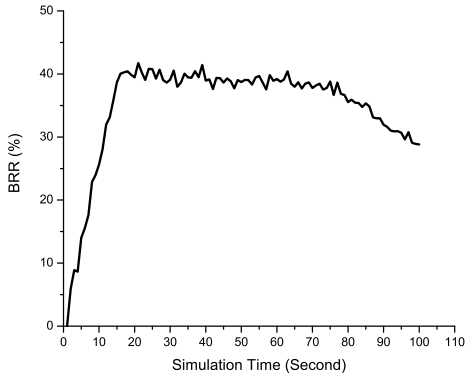


Figure 3.9 Simulation results of *BRR*.

in the system is very light. Therefore, only few connections need to recycle the unused 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. 3.10 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

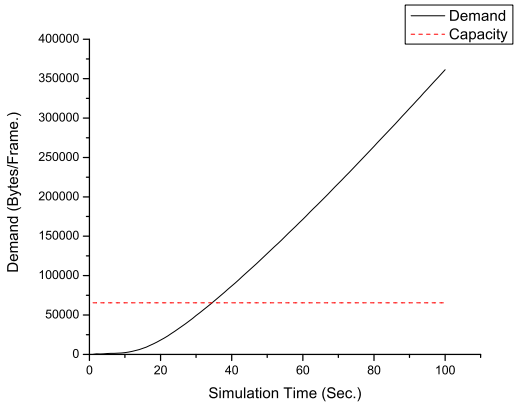


Figure 3.10 Total Bandwidth Demand

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. 3.11 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 Stage 1 and 2 described in section 3.5 that there is no significant improvement on our scheme when the network load is light. As the traffic increases, the TG reaches around 15 to 20%. It is worth to note that the TG reaches around 20% at 35th second of the simulation time. It matches the time that the bandwidth demand reaches the system capacity shown in Fig. 3.10. Again, it confirms our early observation (Stage 3 and 4 in section 3.5) 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% 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 can

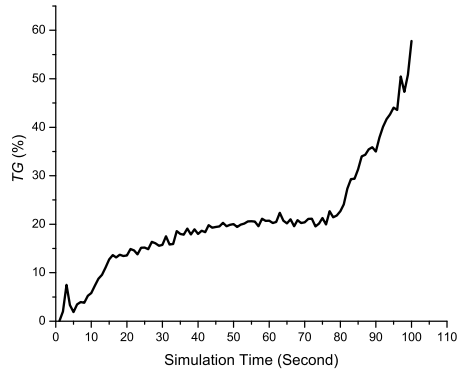


Figure 3.11 Simulation results of TG

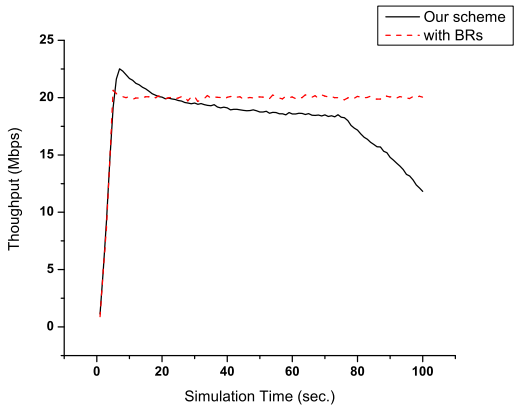


Figure 3.12 Comparison with the case with BRs

not 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 stage 1 to 4 in Section 3.5.

Fig. 3.12 shows the throughput comparison between our scheme and Case with BRs defined in Section 3.5.6. From the figure, we can obtain that the throughput of 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 3.7.

3.6.4 Theoretical Analysis V.S. Simulation Results

In this subsection, 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 3.2. 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%. Moreover, the theoretical result calculated by equation (3.5) is about 35.29%. 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 equation (3.20), the theoretical percentage of RMs coverage is from 42 to 58%. Additionally, the result from our simulation is 48.7% 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 can not be recycled: 1) CSs cannot receive RMs sent by their corresponding TSs. 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% which includes both factors described above. By doing further investigation, we find that about 51.3% of failures is because the CS cannot receive a RM form the corresponding TS. The rest of failures, about 10.2%, are caused by no data to be transmitted while the CS receives a RM. Based on this observation, three scheduling algorithms are proposed in Section 3.7 to mitigate the affection of these factors for improving the recycling performance.

3.7 Further Enhancement

As our investigation, one of the factors causing recycling failures is that the CS does not have data to transmit while receiving a 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 4.

Algorithm 4 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 3, 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 *history-Based Scheduling Algorithm* (HBA), is proposed. The HBA is summarized in Algorithm 5. 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 a 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 a SS which is not on the BL of the TS as the CS. After collecting enough history, the BL of each TS should contains 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 a SS which is able to receive the RM as the CS.

Algorithm 5 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. 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 6

Algorithm 6 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 can not 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.

3.8 Simulation results of enhancement

The simulation model for evaluating these scheduling algorithms is same as the model presented in section 3.6. 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 5 connections. The simulation results of *TG* is shown in Fig. 3.13. Before the 15th second of simulation time, the *TG* may be negative. It means 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 3 can achieve averagely 20% of throughput. The RBRFA can further improve the throughput to 26% 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%. 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 3.6 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% improvement) since it combines both advantages of HBA and RBRFA.

The comparison of BRR is shown in Fig. 3.14. The results consist with the results of *TG* shown above. The HSA has the highest *BBR*. Moreover, the HBA achieves the higher *BBR* than the RFA does. Additionally, it is worth noting that the *BRR* of the RRFA can not be more than 50% even when the network is fully loaded. It is because, based on our investigation in section 3.6, there is only 48.7% of probability that a CS can receive a RM

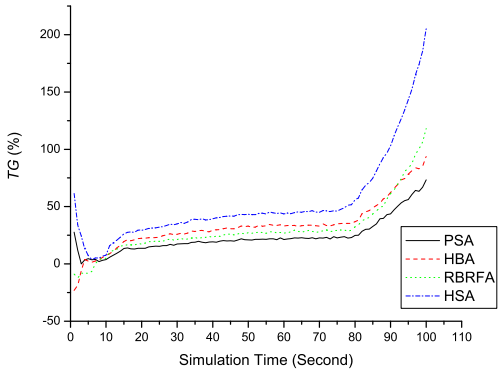


Figure 3.13 Simulation results of TG among all scheduling algorithms

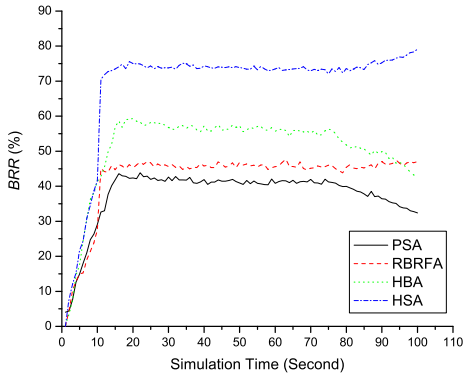


Figure 3.14 Simulation results of BBR among all scheduling algorithms

successfully.

The comparison of the total bandwidth demand is shown in Fig 3.15. 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. 3.16. Similar to the simulation results of TG and BRR. The HSA has the best improvement on delay due to the highest throughput it achieves.

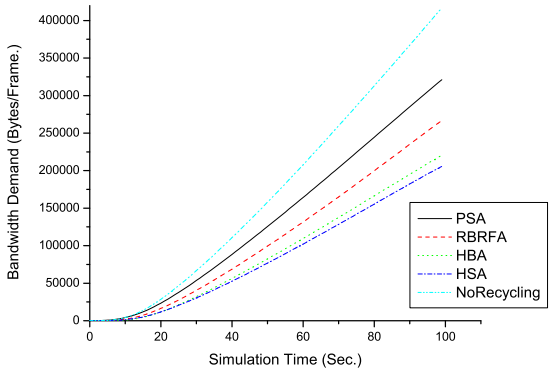


Figure 3.15 Simulation results of bandwidth demand

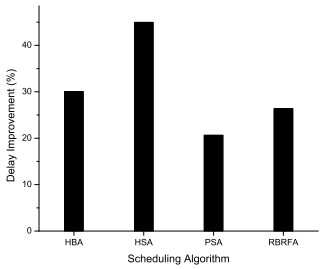


Figure 3.16 Simulation results of delay improvement

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

CHAPTER 4. Design and Analysis of Bandwidth Reservation Game in IEEE 802.16 Networks

A paper to be submitted to IEEE Transactions on Mobile Computing

David Chuck and J. Morris Chang

Abstract

Due to exclusive usage of bandwidth reservation in IEEE 802.16 networks, bandwidth may not be utilized efficiently all the time when the reservation is greater than the bandwidth demand of subscriber stations (SSs). In this paper, we aim to help the SS make the optimal bandwidth reservation such that the overall system bandwidth utilization is maximized while satisfying QoS requirements. We investigate a centralized scheme that the base station (BS) has the completed traffic information of each SS. We further propose a bandwidth reservation (BR) game to help the SS make its bandwidth reservation. Each SS focuses on maximizing its payoff calculated by the utility function. In our utility function, we consider both QoS requirements and total bandwidth demand in the network (TBD) and aim to maximize the system bandwidth utilization while satisfying QoS requirements of each SS. Due to different QoS requirements, the utility function is customized for each scheduling class. The existence and uniqueness of Bayesian Nash equilibrium are demonstrated. In our numerical analysis, we obtain the optimal solution for the centralized scheme through AMPL and investigate the price of anarchy of the proposed game. Our numerical and simulation results show that the network utilization achieved by the proposed game is very close to optimal solution.

4.1 Introduction

The latest IEEE 802.16 standard is one of critical wireless medium access technologies for the forth generation (4G) network(52). One of the fundamental features in IEEE 802.16 networks is to provide quality of service (QoS) guaranteed services. Radio resource reservation is employed in the IEEE 802.16 standard to achieve this feature. In order to serve a wide variety of applications, all applications from upper layer are mapped into connections. Each connection is classified into different types of scheduling class depending on the QoS requirements of applications. A request/grant bandwidth allocation mechanism is specified in the IEEE 802.16 standard. Each subscriber station (SS) requests the required bandwidth from the base station (BS) via bandwidth requests to satisfy the QoS requirements of connection. After receiving a request, the BS makes scheduling decisions to determine the bandwidth allocation for each SS. The SS has exclusive privilege to utilize this allocated bandwidth. However, due to the nature of variable bit rate (VBR) applications, it is very challenging for the SS to make the optimal bandwidth reservation. An inappropriate bandwidth reservation may expose the connection under the risk of the failure to satisfy QoS requirements because of the insufficient bandwidth allocation, or degradation of system performance due to over-requesting bandwidth. In the paper, we focus on the bandwidth reservation problem and aim to provide a solution to help the SS make optimal reservation such that overall system bandwidth utilization is maximized while providing QoS guaranteed services.

The system bandwidth utilization is degraded when the SS over-requests bandwidth. However, the level of degradation depends on the total bandwidth demand in the network, denoted as TBD in this paper. TBD is defined as the summation of the desired bandwidth reservation for each SS. It might be larger than the bandwidth capacity supported by the BS in a heavily loaded network. It is worth noting that TBD only indicates the network traffic demand and may not equal to the total allocated bandwidth. When TBD is small, the SS shall request more bandwidth to reduce data latency. The available bandwidth gets depleted

when TBD increases. At this time, the SS should conservatively make minimum bandwidth reservation such that QoS guaranteed service can be provided to all SSs. Consequently, the SS faces a critical choice of making optimal bandwidth reservation for its connections and both factors (i.e., QoS requirements and TBD) should be considered while making reservation.

Although bandwidth requests have been defined in the standard for bandwidth reservation, however, any specific bandwidth request-grant algorithms are not standardized so that proprietary implementations may be used by the equipment vendors. There are several scheduling frameworks and scheduling algorithms proposed in the literature (54)−(59). However, many of them focus on QoS architecture and scheduling algorithms in the BS to satisfy the diversity of QoS requirements. In (60), Park *et al.* proposed a distributed dual feedback bandwidth request algorithm operated and aimed to optimize the bandwidth usage such that the bandwidth utilization is maximized. However, they only considered QoS requirements of connections. Thus, each connection always requests the minimum amount of bandwidth to *just* satisfy the QoS requirements even when TBD is low.

In our previous work (61), a mechanism, named *Bandwidth Recycling*, is proposed to passively recycle the unused bandwidth such that the system bandwidth utilization is improved. However, this mechanism may not be able to recycle all unused bandwidth due to possible failures of recycling. In this paper, we focus on an active bandwidth allocation approach and aim to *maximize the system bandwidth utilization while maintaining QoS requirements under varied* TBD. Here, the bandwidth described in this paper is in terms of bytes per second. We narrow our focus to point-to-multipoint (PMP) mode in which transmissions only exist between BS and SS. We first assume that the BS has all traffic information (e.g., queue status) at each SS. The BS performs centralized scheduling and achieves optimal bandwidth allocation for all SSs. However, in order to perform optimal allocation, the BS needs to have updated traffic information from SSs very frequently. It may require extensive amount of message exchanges between BS and SS. This approach is

named *centralized scheme* in this paper.

To avoid additional message exchanges, we further propose a distributed bandwidth reservation (BR) game to help SSs determine bandwidth reservation and compare the numerical results between centralized scheme and BR game. In BR game, each SS determines its desired bandwidth reservation independently. It does not require message exchanges in the centralized scheme. However, due to independent operations, it is very challenging for a SS to gather the completed bandwidth reservation information and scheduling class of each connection in other SSs. The SS may need to "guess" the information of bandwidth reservation and scheduling class of connections running in other SSs. Therefore, BR game is categorized as an incomplete information game or a *Bayesian game* (62). Additionally, each SS is self-interested in requesting bandwidth as much as possible such that the queue size of its connection is minimized. It makes the cooperative behavior, such as cooperation of maximizing network utilization, hard to achieve. Therefore, this game is classified as an non-cooperative game.

In the formulation of BR game, all SSs are modeled as players in the game and assumed to act rational of maximizing the payoff of each connection. The payoff is computed by two indexes: *satisfaction index* (SI) and *penalty index* (PI). SI represents QoS satisfaction of the connection. PI, on the other hand, stands for the cost of bandwidth when the SS makes a bandwidth reservation. It has an negative correlation with the reputation of connection. A better reputation leads to a lower PI. The reputation depends on the bandwidth utilization of connection for the allocated bandwidth. The connection maintaining good bandwidth utilization earns good reputation. As our objective of maximizing overall bandwidth utilization, each SS tries to maximize its payoff such that the bandwidth utilization is maximized while satisfying the QoS requirements. The objective is achieved when every SS reaches its maximum payoff. The BS in this game plays a role of enforcing the penalty. It is worth noting that the reason of enforcing the penalty at the BS is due to the operation of IEEE 802.16 networks. The BS does not have objective to achieve and is not a

player in the proposed game, either. The SS has the incentive to maintain a good reputation for its connection in order to minimize the penalty such that the payoff is maximized.

In addition to reputation, TBD is another factor determining the value of PI. When the network is heavily loaded, the cost of bandwidth should be high. Consequently, PI grows fast when TBD is large. In our design, the payoff received by the SS is the difference between SI and PI (i.e., $SI - PI$). To maximize the payoff,the SS focuses on determining the amount of requested bandwidth such that the difference between SI and PI (i.e., SI−PI) is maximized.

In our numerical analysis and simulation, both centralized scheme and BR game are evaluated. We first obtain the optimal solution of centralized scheme through **A** **M**athematical **P**rogramming **L**anguage (AMPL), an algebraic modeling language for describing and solving high-complexity problems for large-scale mathematical computation. Further, we implement the proposed game theoretic framework among BS and SS in a simulator. From the simulation, we measure the overall system utilization in the game. We compare it to the optimal solution obtained from the centralized scheme to investigate the price of anarchy of the game. The price of anarchy is defined as the difference between the optimal solution and the worst case of Bayesian Nash equilibrium (BNE) solution. We evaluate the overall system utilization as well as throughput for connection in each scheduling class under different TBDs. Our numerical and simulation results show that the system utilization achieved by the BR game is close to the optimal bandwidth allocation and QoS requirements of connection in all scheduling classes can be satisfied.

In summary, the contributions of this paper are as follows:

- A centralized scheme is formulated to obtain the optimal bandwidth allocation.
- We formulate a distributed BR game such that the overall system bandwidth utilization is maximized while providing QoS guaranteed services.
- We define the utility function comprising SI and PI, representing the QoS satisfaction of each connection and the cost of bandwidth when the SS makes bandwidth reser-

vation, respectively. Due to different QoS requirements and traffic characteristics, we customize the utility function for each scheduling class.

- We investigate the existence and uniqueness of BNE
- The numerical analysis evaluates both centralized scheme and BR game. Our results show that the network utilization for each scheduling class achieved by the BR game is close to the optimal solution while providing QoS guaranteed services.

The rest of paper is organized as follows. An overview of the IEEE 802.16 standard is presented in Section 4.2. The related works of applying game theory in wireless networks are in Section 4.3. In Section 4.4, we introduce the network model and then present the centralized scheme in Section 4.5. The BR game formulation is presented in Section 4.6. The utility function of each scheduling class is shown in Section 4.7. In Section 4.8, we present the investigation of BNE. At the end, the numerical and simulation results and and conclusion are given in Section 4.9 and 4.10, respectively.

4.2 Bandwidth Reservation in IEEE 802.16 Networks

IEEE 802.16 standard is one of the most promising wireless medium access technologies in 4G networks. Relying on bandwidth reservation, IEEE 802.16 networks are able to support QoS guaranteed services. When a SS starts to serve a new application, it has to map the service flow of this application into a connection with one of scheduling classes and a set of QoS parameters (e.g., MSR) based on the QoS requirements of the application and the subscription level of the SS. After this, the SS starts to make an admission control request for this connection. After receiving this request, the BS makes the decision of admitting or rejecting this request based on the admission control policy and the available bandwidth. This process helps the SS identify the guaranteed bandwidth for this connection. It is worth noting that this guaranteed bandwidth is not allocated to the SS instantly. A bandwidth-on-demand scheme is employed in IEEE 802.16 networks. Initially, the SS has

no bandwidth allocated and it request bandwidth from the BS based on the actual traffic demand via the request/grant mechanism defined in the IEEE 802.16 standard. The SS encapsulates the desired amount of bandwidth in a bandwidth request and transmits it to the BS. After receiving this request, the BS performs bandwidth allocation to each SS based on bandwidth availability and scheduling policies. To support a wide variety of applications, all traffic is classified into one of five scheduling classes defined in the IEEE 802.16 standard based on QoS requirement: Unsolicited Grant Service (UGS), Real Time Polling Service (rtPS), Extended Real Time Polling Service (ertPS), Non-real Time Polling Service (nrtPS), Best Effort (BE). The characteristics for each scheduling class can be found in (1) and (53).

All transmissions between BS and SS are relied on unidirectional connections associating to service flows characterized by a set of QoS parameters: *maximum traffic rate, minimum sustained rate (MSR) and maximum tolerable delay.* Transmissions are classified into downlink (DL) and uplink (UL) transmissions based on the transmission direction. DL transmissions are transmissions from BS to SS and UL transmissions are in the opposite direction. Because BS always has completed traffic information for all DL connections, the optimal scheduling decision can be made easily. Unfortunately, it is challenging for the BS to gather the completed information to perform scheduling decisions for UL connections. The BS has to rely on the information provided by the SS via bandwidth requests to perform UL bandwidth allocation.

4.3 Related Game theoretic works for wireless networks

Game theoretic approaches have been widely used in wireless networks for designing mechanisms to reach an equilibria by modeling both benefit and cost of a wireless node. There has been growing interest in applying game theory to several popular research topics in wireless networks such as admission control, power conservation and resource allocation. A good tutorial (63) written by Felegyhazi and Hubaux introduces the basic concepts and strategies of the game theory in wireless networks.

In (64), Niyato *et. al* proposed an admission control mechanism based on game theory for IEEE 802.16 networks. The players in the game are newly arrived connections and the BS. By modeling the QoS satisfaction to the amount of corresponding allocated bandwidth as the utilities, Nash equilibrium was proposed as a solution of determining whether the newly arrival connection is admitted or not. In CDMA system, the authors in (65) propose a similar approach to determine whether the network admits or rejects the newly arrival user such that the resource utility is maximized. Instead of considering one resource provider, the authors in (66) focus on the admission control among multiple resource providers. The goal of this game is to produce an integrated pricing and admission control policy that achieves the network provider optimum utility, while ensuring the satisfaction of all sides.

Many game theoretic approaches have been proposed to solve the power conservation problem. In (67), a framework for power control in sensor networks has been proposed to find the optimal threshold of transmission power for each node. By considering the limited power of each sensor, the authors focused on finding the minimum power threshold such that the utility of each sensor is maximized. In addition to the continuous power levels, the authors determine the number of power levels based on the probability density function of interference to minimize the distortion factor which is defined as the difference between the best possible utility obtainable with continuous power level and the best possible utility obtained with the number of discrete power levels. Niyato *et. al* proposed a non-cooperative game theoretic technique (68) to investigate energy harvesting technologies for autonomous sensor networks. The authors went through the related works on energy efficiency for sensor networks using energy harvesting technologies. At the end, Nash equilibrium was proposed to determine the optimal probabilities of sleep and wakeup states for energy conservation. A seller-buyer game for cooperative communications is proposed in (69). A two-level Stackelberg game is employed to jointly consider the benefits of the source node (modeled as a buyer) and the relay nodes (modeled as a seller). The objective of this Stackelberg game aims to not only help the source find a relay node at a relatively better

location and use the optimal amount of power to communicate with the relay but also maximize the utilities of the relay node.

An non-cooperative game theoretic approach for dynamic spectrum sharing in cogitative networks was proposed in (70). The authors modeled the problem of spectrum sharing as a seller-buyer game in which primary users sell spectrum opportunities to secondary users and the secondary users adapt the spectrum buying behavior by observing the variations in price and quality of spectrum offered by the different primary users. The Nash equilibrium is considered as the solution of the game in terms of the size of spectrum offered to secondary users and the spectrum price. In (71), the authors modeled the channel allocation problem in multihop wireless networks as a hybrid game in which the game is cooperative within a communication session but non-cooperative among sessions. This game aims to maximize the achieved data rates of communication sessions. In (72), a Bayesian game has been modeled for Network Selection in Heterogeneous Wireless Networks. The objective of this game aims to help user equipment select a type of networks such that load balancing is achieved. The author in (73) proposed a bidding model by applying Bayesian game to help mobile user make vertical hand-off decisions.

In this paper, we model the bandwidth allocation problem in IEEE 802.16 network as an non-cooperative game. Unlike a seller-buyer game modeling BS and SS as seller and buyer, respectively, the players in this game are SSs. Each SS focuses on maximizing its own payoff. The objective of maximizing the system bandwidth utilization with QoS guaranteed service is achieved as the result that all SSs reach their maximum payoff. Because of the operation of IEEE 802.16 networks, the bandwidth allocated to each SS must be assigned by the BS. Thus, the BS in our scheme is responsible for enforcing the penalty to each SS. It does not have an objective to reach. Moreover, due to the implementation of PMP mode, there is no message exchange between SSs. It may be challenging for each SS to gather the traffic information (e.g., scheduling class of connections) of other SSs. It makes the proposed game as an incomplete information game or Bayesian game. In summary, the proposed game is

modeled as an non-cooperative Bayesian game with the objective of maximizing the network utilization while providing QoS guaranteed services.

4.4 System Model

We consider a network comprising a BS and $|N|$ SSs. The BS is located at the center of the geographical area and the $|N|$ SSs are randomly distributed in its service coverage. We use $N = \{1, 2, \ldots, n\}$ to denote the set of SSs in the network. For simplicity, we assume that each SS serves one connection randomly classified into one type of scheduling classes except UGS due to the unadjustable bandwidth allocation for UGS connections. Furthermore, according to the specification of ertPS in the IEEE 802.16e standard, the behavior of requesting bandwidth for an ertPS connection is same as the one of a rtPS connection. Consequently, we further narrow our interest into three types of scheduling classes, rtPS, nrtPS and BE, and represent them in a set $T = \{rt, nrt, be\}$.

According to IEEE 802.16 standard, the SS may communicate with the BS via different types of modulation. Again, the bandwidth throughout the paper is presented in terms of bytes per second. Furthermore, according to the IEEE 802.16 standard, the bandwidth requested by each SS is also represented in bytes. The BS knows the modulation used by each SS. Consequently, the issue that the different types of modulation used by each SS can be considered by the BS when the BS makes decisions for bandwidth allocation.

Each connection shall pass the admission control before being served by a SS. According to our admission control policy, the BS shall guarantee bandwidth to provide the MSR to each connection. Additionally, according to the IEEE 802.16 standard, both rtPS and nrtPS connections require non-zero MSR. However, it is not necessary for BE connections due to flexible QoS requirements. In our system, we assume non-zero MSR for both rtPS and nrtPS connections and zero MSR for BE connections.

In each frame, the SS determines the amount of required bandwidth for the next frame. If it is different from the one allocated in the current frame, the SS specifies the difference

in the extended piggyback request (EPBR) field of grant management subheader. The size of EPBR field is 11 bits representing two operation modes. If the most significant bit is set to zero, then this request is an incremental request. The rest of 10 bits denote the increment of bandwidth. Otherwise, this request is an aggregate request in which the rest of 10 bits stand for the amount of requested bandwidth for the next coming frame. The SS may not transmit a bandwidth request if the amount of required bandwidth for the next frame is same as the one allocated in the current frame. Since it is possible that no bandwidth is allocated to BE connections, bandwidth requests may not be able to transmit via piggyback. We assume that BE connections can always transmit bandwidth requests through contention resolution if no bandwidth is allocated.

4.5 Bandwidth Allocation with Complete Information

In this scenario, we assume that the BS has the completed traffic information of the connection served by each SS. We formalize this bandwidth allocation as the following linear programming problem:

$$\max \frac{1}{B_T} \sum_{\forall i \in N} x_i \tag{4.1}$$

such that

$$\sum_{\forall i \in N} x_i \leq B_T \tag{4.2}$$

$$x_j \geq \frac{Q_j}{D_j^{max}} \qquad \forall j \in \{rt\} \tag{4.3}$$

$$x_i \geq MSR_i \qquad \forall i \in N \tag{4.4}$$

$$x_i \leq Q_i \qquad \forall i \in N \tag{4.5}$$

The objective function shown in (4.1) is to maximize total bandwidth utilization in the system. x_i and B_T stand for the amount of bandwidth allocated to SS i and the total bandwidth that the BS can support, respectively. Since the allocation is in per frame duration which is a constant in WiMAX,, x_i and B_T can be represented in terms of bytes.

Furthermore, the modulation used for each SS has to be fixed within a frame. Therefore, we can assume B_T is a constant during the frame when the proposed game is performed..

The constraints of this linear programming problem are listed as (4.2)-(4.5). The first constraint shown in (4.2) indicates that the total allocated bandwidth cannot exceed B_T. Formula (4.3) specifies the delay constraint for the SS serving rtPS connections. Q_j and D_j^{max} are the expected amount of queued data and the maximum tolerated delay for a rtPS connection j, respectively. Consequently, the value of $\frac{Q_j}{D_j^{max}}$ indicates the minimum amount of bandwidth requirement for j in order to satisfy its maximum delay requirement. Therefore, formula (4.3) can represent the QoS requirement. The admission control requirement stated in (4.4) represents that the amount of allocated bandwidth cannot be less than the minimum bandwidth requirement. MSR_i is the minimum bandwidth requirement that claimed in admission control process. The last constraint shown in (4.5) ensures that the amount of bandwidth does not larger than the expected amount of queued data for all SSs to avoid the bandwidth wastage.

Although this approach can lead us to the optimal solution, the BS may require to have updated information of queue status from each SS very frequently. It may require large amount of message exchanges which introduce extensive network overhead. To avoid this, we further propose a game theoretic framework to help the SS determine the optimal bandwidth reservation in fully distributed fashion. We also compare the numerical results between this centralized scheme and game. The details of this framework are presented in the following sections.

4.6 Bandwidth Reservation Game

In this section, we first present an overview of the proposed game including motivation, objective and game classification. Then, we introduce the game formulation.

4.6.1 Overview

Although the centralized scheme can lead us to optimal solutions, this scheme is based on the assumption that the BS has the completed traffic information of each SS. To achieve this assumption, the BS may rely on a large number of message exchanges with SSs to gather the latest queuing information from SSs in real time. Furthermore, the decision of bandwidth allocation made by the BS is based on the requested bandwidth claimed by each SS. There is no mechanism to help the BS verify whether this request matches the actual bandwidth requirement of the SS. A mismatched request might degrade the overall bandwidth utilization. Consequently, the motivation of the proposed game is to design a scheme in fully distributed fashion without introducing additional message exchanges between SS and BS while minimizing the gap between allocated bandwidth and real usage. Furthermore, same as the centralized scheme, the objective of the proposed game aims to maximize the overall bandwidth utilization while providing QoS guaranteed service.

In our game, each SS is assumed to act rationally and tends to request as much bandwidth as possible in order to provide the best service to its connection. However, the bandwidth capacity provided by the BS is limited. The SS may compete with each other on bandwidth reservation. It makes cooperative behavior between SSs impossible. Consequently, the proposed game is classified as an non-cooperative game. Furthermore, no communication is allowed between SSs in PMP mode. It makes SS difficult to gather traffic information (e.g., scheduling class of connection) of other SSs. Thus, the proposed game is categorized as an incomplete information game or Bayesian game. In summary, the proposed game is classified as an non-cooperative Bayesian game.

In the proposed game, each SS focuses on requesting its own bandwidth such that its payoff is maximized. The payoff is calculated by the utility function which will be presented in Section 4.7. Due to the independent operation of each SS, it is possible that the summation of bandwidth requested by each SS exceeds the total bandwidth supported by the BS. Although each SS specifies the amount of requested bandwidth, the actual

bandwidth allocation is made by the BS due to the operation of a IEEE 802.16 network. The BS has the obligation to ensure that the total allocated bandwidth does not exceed the bandwidth capacity. In this paper, we assume that the BS grants bandwidth partially based on the weight of requested bandwidth when total requested bandwidth is more than the bandwidth capacity of BS.

4.6.2 Game Formulation

The players in the proposed game are SSs. When the SS starts to serve a new application, this application is mapped into a connection with the scheduling class corresponding to its QoS requirements. In this paper, we consider three types of applications: video streaming, FTP and web browsing. We assume that each connection serves one of these types of applications in uniformly distributed fashion. We use ρ^{rt}, ρ^{nrt} and ρ^{be} to represent the probabilities that a connection serving video streaming, FTP and web browsing, respectively. Based on the QoS requirements of each type of applications, the connection serving video streaming, FTP and web browsing is mapped as rtPS, nrtPS and BE connections, respectively. Moreover, the connection can be mapped to only one scheduling class and this scheduling class cannot be changed until the connection is terminated. The amount of bandwidth requested for this connection is denoted as $b \in [B^{min}, B^{max}]$, where B^{min} and B^{max} are the minimum and maximum amount of bandwidth that can be requested for this connection, respectively.

The bandwidth supported by the BS is shared among all SSs. Each SS is assumed to act rationally and tries to request bandwidth as much as possible to minimize the queue length of each connection in order to provide the best service quality to its connections. Consequently, cooperative behavior between SSs may not be possible and this leads the proposed game as an non-cooperative game. Moreover, PMP mode is assumed. There are no communications between SSs. It is very challenging for a SS to gather the information of scheduling class and queue status of other SSs. Therefore, we model this N-SS bandwidth

allocation problem as an incompleted information game or Bayesian game. Each SS in the proposed game only know the traffic information of its own connection. Therefore, the SS needs to predict the necessary information of other SSs (e.g., scheduling class of connections) in order to determine its bandwidth reservation.

For any SS i, b_i denotes the amount of bandwidth requested for its connection. When the SS selects its strategy, it needs to have a belief on the scheduling class of other SSs and forecast the strategies selected by other SSs based on the belief. We call the set of forecasted strategies as *deleted strategy profile* denoted as $b_{-i} = \{b_1, \ldots, b_{i-1}, b_{i+1}, \ldots, b_n\}$. In the Bayesian game, the SS needs to predict each element in the deleted strategy profile based on the type (i.e., scheduling class of connection) of other SSs. Suppose b_k represents any strategy for SS k in b_{-i}. We can represent b_k as:

$$b_k = \sum_{t \in \{rt, nrt, be\}} \rho_k^t \cdot b_k^t \tag{4.6}$$

where b_k^t is the strategy of SS k that SS i predicts for the SS k in scheduling class t. The payoff of i is calculated by the utility function denoted as u_i which will be defined in the later section. In summary, the *N-SS* Bayesian game can be completely characterized as:

- Player set: $N = \{1, 2, \ldots, n\}$
- Type set: $\Upsilon = t_1 \times t_2 \times \ldots \times t_n$, where $\times$ stands for the Cartesian product and $t_i \in T = \{rt, nrt, be\}$, $\forall i \in N$.
- Action set: $B = b_1 \times b_2 \times \ldots \times b_n$.
- Probability set: $\Omega = \{\omega_1, \omega_2, \ldots, \omega_n\}$, where $\omega_i \in \{\rho_i^{rt}, \rho_i^{nrt}, \rho_i^{be}\}$
- Utility function set: $U = \{u_1, u_2, \ldots, u_n\}$.

4.7 Utility Functions

First, we present the general formulation of utility function. Because of different QoS requirements and traffic characteristics, the utility function is customized for each scheduling

class. The details are presented in the following subsections.

4.7.1 General Formulation

When the SS determines the amount of requested bandwidth for its connection, two factors should be considered: QoS requirements and TBD. Satisfying QoS requirements ensures the connection to have enough bandwidth for successful operation. TBD allows the SS to adjust the bandwidth reservation corresponding to the current traffic load in the network. Each SS is assumed to act rationally and focuses on providing the best service to its connection. However, due to limited bandwidth capacity, a "benefit-cost" concept is adopted to regulate the bandwidth request of each SS.

In our design, the utility function comprises two indexes: satisfaction index (SI) and penalty index (PI). SI represents the service quality satisfaction of the connection corresponding to the amount of requested bandwidth. PI stands for the current cost to request bandwidth. The cost is determined by TBD as well as the reputation of the connection, where the reputation represents the bandwidth utilization of the connection for the allocated bandwidth. The network with high TBD has high cost for requesting bandwidth. This makes SS request less bandwidth to just satisfy the QoS requirements. On the other hand, the SS can requests more bandwidth when TBD is low. The reputation represents the efficiency of bandwidth reservation utilized by the connection. Due to the operation of IEEE 802.16 networks, all bandwidth should be allocated by the BS. Therefore, in our game, the BS is responsible to enforce the penalty by allocating the bandwidth based on the history of bandwidth utilization of the SS as well as TBD. It is worth noting that the BS only performs penalty enforcement and does not involve in the game operation. The objective of the proposed game is to maximizing the system utilization. It is expected to achieve by each SS with the maximum payoff of its connections. The payoff is defined as $(SI - PI)$. The detail will be presented later in this section.

Since SI is only related to the amount of requested bandwidth, we model it as a function

of the requested bandwidth. On the other hand, PI is associated with two factors (i.e., the bandwidth utilization of connection and TBD). Therefore, we design it with two sub-indexes: bandwidth demand sub-index (BDI) and performance sub-index (PSI). BDI indicates the bandwidth demand in the network. It is positively correlated to TBD which is the outcome of strategies selected by each SS. It is formed as a function of not only the strategy selected by the SS but also the strategies selected by other SSs. Thus, for any SS i, the TBD for this SS can be expressed as:

$$TBD(b_i, b_{-i}) = \frac{b_i + \sum_{b_q \in b_{-i}} b_q}{B_T} \tag{4.7}$$

Note that TBD may be larger than 1 indicating that the network is heavily loaded. Again, B_T is the bandwidth capacity supported by the BS. Consequently, BDI for SS i can be represented as $BDI_i(TBD(b_i, b_{-i}))$.

PSI measures the bandwidth utilization of the connection and can be formed as a function of the strategy selected by the SS. Unlike BDI, PSI is negatively correlated to the bandwidth utilization of the connection. It brings smaller value of PSI for a connection with good utilization. With these two sub-indexes (i.e., BDI and PSI), the PI for SS i is designed as:

$$PI_i(b_i, b_{-i}) = BDI_i(TBD(b_i, b_{-i})) \cdot PSI_i(b_i) \tag{4.8}$$

When the network is lightly loaded (i.e., TBD is small), it indicates that there is more available bandwidth in the network. It makes BDI smaller. and the SS can take the advantage to request more bandwidth to shorten the data latency. However, when there is less available bandwidth in the network, BDI gets large and increases the cost of bandwidth request (i.e., PI). The SS tends to well-utilize the requested bandwidth to minimize PSI such that the PI is minimized. This design gives the SS flexibility to adjust its bandwidth reservation depending on the total network demand in the network.

A "benefit-cost" concept is adopted in the utility function. SI representing the satisfaction of service quality for the connection is considered as the benefit corresponding to the amount of requested bandwidth. On the other hand, PI is the cost. The payoff is the net

benefit that the SS receives. It is defined as the difference between benefit and cost (i.e., $SI - PI$). Additionally, it is worth noting that according to the IEEE 802.16 standard, the scheduling class of a connection cannot be changed after the creation of the connection. It makes that only one possible type for the player. However, the player still needs to forecast all possible scheduling classes for all other players. Thus, for any SS i, the payoff can be written formally as:

$$u_i(b_i, b_{-i}) = SI_i(b_i) - PI_i(b_i, b_{-i}) \tag{4.9}$$

Each SS tries to determine the amount of bandwidth to be requested such that the payoff is maximized. It is equivalent to find the amount of requested bandwidth such that the value of SI is maximized while minimizing the value of PI.

Due to different QoS requirements, we customize the utility function for each scheduling class presented in the following subsections. We first introduce the utility function for rtPS connections followed by the one for nrtPS and BE connections.

4.7.2 rtPS

4.7.2.1 SI

The traffic classified as a rtPS connection is usually delay-sensitive. Data arriving from the upper layer need to be transmitted within a limited period of time. This period of time is usually referred to the maximum delay requirement. Generally speaking, the data start to accumulate in queue when the amount of reserved bandwidth is less than the mean data generation rate. Therefore, the minimum bandwidth requirement should match to the mean data generation rate to avoid the overflow problem. This minimum bandwidth requirement should be passed to the BS as MSR during admission control procedure in order to have guaranteed bandwidth for maintaining QoS requirements.

The QoS satisfaction of a rtPS connection drops significantly if the incoming data start to accumulate in queue. The dropping rate decreases when less amount of bandwidth is allocated since the service quality is too bad to be recovered. On the other hand, the SS has

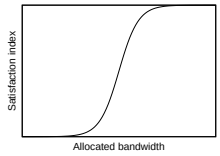


Figure 4.1 A sample of sigmoidal-like function

more capacity to support unexpected burst data arrival or jitter when the amount of reserved bandwidth is larger than the MSR. Thus, the QoS satisfaction increases significantly when the bandwidth is larger than the MSR. The increasing rate drops with the increase of the bandwidth reservation due to less importance for the connection at this time.

To model QoS satisfaction of real time traffic, sigmoid function has been used widely in the literature (75)-(77). A sample of sigmoid function is shown in Fig. 4.1. We observe that the sigmoid function is partially concave and partially convex. This feature matches the traffic characteristics of real time traffic described above. The convex part is used to model the growing of SI when the amount of reserved bandwidth is less than MSR. The concave part represents the change of SI after MSR is reached. The infection point models the status that the amount of reserved bandwidth matches to the MSR of the connection. Consequently, the SI for a rtPS connection running on SS i can be represented as:

$$SI_i^{rt}(b_i) = \frac{1}{1 + e^{-k_i^{rt}(b_i - MSR_i)}} \tag{4.10}$$

where $(b_i - MSR_i)$ represents the rate of data accumulation for the connection. When $(b_i - MSR_i) < 0$, the data arriving from the upper layer are accumulated in queue. The smaller $(b_i - MSR_i)$ leads to the queue to be built up faster. k_i^{rt} represents the sensitivity of SI to data accumulation in queue, where $k_i^{rt} > 0$. When k_i^{rt} is larger, SI_i^{rt} is more sensitive to the data accumulation rate.

4.7.2.2 PI

As shown in equation (4.8), PI comprises two sub-indexes: BDI and PSI, representing TBD and bandwidth utilization of connection, respectively. BDI is a function of TBD

and can be represented for SS i as:

$$BDI_i^{rt}(b_i, b_{-i}) = \left(TBD(b_i, b_{-i})\right)^{\beta_i^{rt}} \tag{4.11}$$

where β_i^{rt} is the parameter for the rtPS connection running on SS i. It describes the sensitivity of PI to TBD.

PSI relates to the bandwidth utilization of a connection. This can be estimated based on the expected amount of data stored in queue. This amount of data can be predicted by the data stored in queue (denoted by Q^{in}) plus the expected amount of data arriving within this frame (denoted by Q^f). The SS has to ensure that the maximum delay requirement can be satisfied. Thus, when the maximum delay requirement is not satisfied, PSI should be very small to minimize the value of PI such that the SS has opportunities to request more bandwidth. After the maximum delay requirement is satisfied, the PSI becomes large to represent the need of bandwidth. Thus, the PSI of connection i can be represented as:

$$PSI_i^{rt} = \left(\frac{b_i \cdot t_i^{max}}{Q_i^{in} + Q_i^f}\right)^{\varphi_i^{rt}} \tag{4.12}$$

where t_i^{max} is the maximum delay requirement of the connection served by SS i. We consider t_i^{max} to accommodate bursty data such that the maximum delay requirement can be achieved for rtPS connections.. Similar to BDI, φ_i^{rt} describes the sensitivity of PI to the bandwidth utilization.

In our admission control policy, the BS should provide the guaranteed bandwidth to ensure that the MSR can be provided. It makes no penalty to the connection before the amount requested bandwidth reaches its MSR. Thus, PI should be zero when the amount of reserved bandwidth is less than MSR and start to grow hereafter. In summary, PI of the rtPS connection running on SS i is presented as:

$$PI_i^{rt}(b_i, b_{-i}) = \begin{cases} BDI_i^{rt}(b_i, b_{-i}) \cdot PSI_i^{rt}(b_i), & b_i \geq MSR_i^{rt} \\ 0, & \text{Otherwise} \end{cases} \tag{4.13}$$

4.7.3 nrtPS and BE

Delay tolerated traffic usually has more flexible QoS requirements comparing to delay sensitive traffic. In IEEE 802.16 standard, nrtPS and BE are designed to serve this type of traffic. The difference between nrtPS and BE is that the MSR of a nrtPS connection cannot be zero. However, the BS can allocate zero bandwidth to BE connections since they do not have any QoS requirements.

4.7.3.1 SI

Elastic functions (75)-(77) are typically used to model the satisfaction of service quality for delay tolerated traffic. This type of functions is always increasing, but the increasing rate descends when the amount of reserved bandwidth increases. It is because the allocated bandwidth becomes less important when there is more bandwidth allocated. We adopt a logarithm function to represent the SI for both nrtPS and BE connections. This logarithm function can be presented as:

$$SI^t(b^t) = \kappa^t \log(1 + h^t b^t) \tag{4.14}$$

where κ^t and h^t are parameters corresponding to each scheduling class $t \in \{nrt, be\}$. These parameters describe the sensitivity of SI to bandwidth reservation. b^t is the amount of requested bandwidth for the connection with the corresponding scheduling class.

4.7.3.2 PI

Similar to a rtPS connection, the PI for both nrtPS and BE connections is calculated by BDI and PSI. NrtPS and BE connections should have more responsibility to maintain high bandwidth utilization due to flexible QoS requirements. Therefore, PSI for both nrtPS and BE connections is designed based on their bandwidth utilization, which can be presented as:

$$PSI^t = \left(U(b^t)^{-1}\right)^{\varphi^t} \tag{4.15}$$

where

$$U(b^t) = \begin{cases} 1 & Q^{in} + Q^f \geq b_t \\ \dfrac{Q^{in} + Q^f}{b^t} & \text{Otherwise} \end{cases} \tag{4.16}$$

$U(b^t)$ represents the bandwidth utilization associated with the reserved bandwidth b^t. Q^{in} and Q^f represent the queued data and the expected data arriving within the current frame, respectively. φ^t is the parameter representing the sensitivity of PSI to the connection performance corresponding to scheduling class $t \in \{nrt, be\}$. PSI is more sensitive to the connection performance when larger φ^t is used.

BDI is based on the outcome of selected strategies by all SSs in the network. Due to flexible QoS requirements of delay tolerated traffic, nrtPS and BE connections should request less bandwidth when the network is heavily loaded. Moreover, although all connections should target to maintain high bandwidth utilization, nrtPS and BE connections should have more responsibility to maintain high bandwidth utilization all the time due to flexible delay requirement. Therefore, unlike rtPS, the PI for nrtPS and BE connections only depends on PSI when the network is lightly loaded. Consequently, BDI for nrtPS and BE connection can be presented as:

$$BDI^t(b^t, b^t_-) = (F^{-t})^{\beta t} \tag{4.17}$$

where

$$F^{-t} = \max\left\{1, TBD(b^t, b^t_-)\right\} \tag{4.18}$$

where β^t are the parameter representing the sensitivity of BDI corresponding to TBD with scheduling class $t \in \{nrt, be\}$. F^{-t} cannot be less than one to ensure that both nrtPS and BE connections maintain high bandwidth utilization when the network is lightly loaded (i.e., TBD is smaller than 1).

As mentioned earlier, the MSR of an nrtPS connection must be nonzero. Moreover, the BS guarantees bandwidth to ensure that the MSR can be provided as our admission control policy. Thus, similar to rtPS connections, the SS should not be penalized before

MSR is satisfied. It makes the PI for nrtPS connections is zero before MSR is reached. However, BE connections do not require any guaranteed bandwidth. Thus, the PI for a BE connection always depends on its bandwidth utilization for the allocated bandwidth. With the consideration of our admission control policy, the PI for nrtPS and BE connections can be summarized as equation (4.19) and (4.20), respectively.

$$PI^{nrt}(b^{nrt}, b_{-}^{nrt}) = \begin{cases} BDI^{nrt}(b^{nrt}, b_{-}^{nrt}) \cdot PSI^{nrt}(b^{nrt}), & b^{nrt} \geq MSR^{nrt} \\ 0, & \text{Otherwise} \end{cases} \tag{4.19}$$

$$PI^{be}(b^{be}, b_{-}^{be}) = BDI^{be}(b^{be}, b_{-}^{be}) \cdot PSI^{be}(b^{be}) \tag{4.20}$$

4.7.4 Discussion

In the subsections above, we introduce the customized utility function for each scheduling class based on QoS requirements and traffic characteristics. The PI of each scheduling class contains two corresponding parameters, β and φ, which represent the sensitivity to TBD and bandwidth utilization of connection, respectively. The larger values of parameters should makes PI more sensitive to the corresponding factor (i.e., TBD or bandwidth utilization of connection). Therefore, it regulates the value of both β and φ must be larger or equal to one. Additionally, the PI which is more sensitive to the factors may lead to unstable bandwidth reservation. Considering the traffic characteristics, the rtPS connection should be able to provide stable bandwidth allocation comparing to nrtPS and BE connections since it needs to constantly satisfying maximum delay requirement. Additionally, BE connections do not have any QoS requirements. They have the lowest priority comparing to nrtPS and rtPS connections. Thus, BE connections should focus more on maintaining their performance as high as possible and reserve very limited bandwidth when the network has a heavy traffic load. In summary, the guidance of parameters between different scheduling

classes can be summarized as:

$$1 \leq \beta^{rt} \leq \beta^{nrt} \leq \beta^{be}$$

$$1 \leq \varphi^{rt} \leq \varphi^{nrt} \leq \varphi^{be}$$

In the proposed game, each SS focuses on requesting its own bandwidth such that its payoff is maximized. The BS shall allocate bandwidth up to the amount of requested bandwidth to achieve the maximum payoff. Although allocating more bandwidth may improve the SI of the SS, it may reduce the bandwidth utilization of the SS and result higher PI. This may hurt the the payoff of the SS.

The BS enforces the penalty based on the product of bandwidth utilization of each SS and the total traffic demand. Due to different QoS requirements and traffic characteristic for each scheduling class, the sensitivity parameter discussed above should be considered in penalty enforcement. Therefore, the guidance of sensitivity parameters between different scheduling classes should be followed while enforcing the penalty and these QoS related parameters are available in the BS..

4.8 Bayesian Nash Equilibrium

4.8.1 Definition

After introducing BR game formulation, what can we expect the outcome of the BR game if every player plays the game rationally and selfishly? Generally, the process of players' decisions usually results in a BNE. In many cases, it stats the "stable" situation after learning and evaluating all players' decisions. It is very important to evaluate such an equilibrium since it represents the performance perdition of an distribution system.

In a more formal definition, a BNE describes a status that no player can benefit more by changing its strategy while other players keep their strategies unchanged. Note that in a strategic form game with completed information, each player focuses on a concrete strategy. However, in a Bayesian game, the player faces to choose a set of strategies, one for each

type that it may encounter. It is also worth noting that the strategy set of a player is independent of the type set of the player. Thus, it is possible that the strategy set is good for all types.

Based on the description above, the BNE in our game can be addressed as follows. Let $u_i(\hat{b}_i, \mathbf{b_{-i}})$ denote the payoff of SS i when player i plays $\hat{b}_i$ and other players play b_j, where $j \neq i$. Thus, the strategy profile for this payoff can be described as:

$$b_1, \ldots, b_{i-1}, \hat{b}_i, b_{i+1}, \ldots, b_n$$

Definition 1 The strategy profile leads to a BNE if $\forall i \in N$ and $b_i \in B$ and for any give b_{-i}, then there exists at least one $b_i^* \in B$ such that

$$u_i(b_i^*, b_{-i}) \geq u_i(b_i, b_{-i})$$

4.8.2 Analysis of Bayesian Nash Equilibrium

It is well known that an equilibrium point may not exist. In the subsection, we are interested in investigating the existence and uniqueness of a BNE in our resource allocation game.

To show the existence of BNE, we need to show that the strategy set of each player is convex, compact and nonempty (79). Moreover, the utility function is concave on the strategy set. The strategy set of each player in our game is nonempty since every admitted connection allows to request bandwidth. Additionally, $b_i \subseteq \mathbf{R}$. Thus, the strategy set is convex and compact. Now we want to show that the utility function is continuous and concave on both b_i and b_{-i}. Our utility function comprises SI and PI. It is easy to show that both SI and PI are continuous on b_i and b_{-i}. Thus, the utility function is continuous on both b_i and b_{-i}. As shown in Section 4.7, PI for all scheduling classes is modeled based on an exponential function. It is clear to conclude that PI is concave for the payoff function. We show the concavity of SI by the following Lemma.

Lemma 1: In the proposed game, the SI for all scheduling classes is concave.

proof: As shown in equation (4.14), the SI for nrtPS and BE connections is modeled by a

logarithm function. Thus, it is clear to conclude that the SI for nrtPS and BE connections is concave. We now focus on the SI for rtPS connections. As shown in equation (4.10), the SI for rtPS connections is modeled by a sigmoid function. Additionally, the infection point of the sigmoid function is based on the MSR of the connection. It separates SI for SS i with a rtPS connection as below:

$$SI_i(b_i): \begin{cases} convex, & 0 \leq b_i \leq MSR_i^{rt} \\ \\ concave, & MSR_i^{rt} \leq b_i < \infty \end{cases}$$

According to our admission control policy, each SS should receive the guaranteed bandwidth until its MSR is reached. Because of this policy, in our design, the PI of all scheduling classes is zero before the MSR is reached. Moreover, the objective of each SS tries to maximize its payoff which is $SI - PI$. Therefore, the SS must request bandwidth which is larger or equal to its MSR. Consequently, we can limit our consideration only to the concave part of the sigmoid function and conclude that the SI for rtPS connections is also concave in the proposed game. In summary, the SI for all scheduling classes is concave in the proposed game.

Based on Lemma 1 and description above, the proof for the existence of BNE is completed. Now, we investigate the uniqueness of BNE. We rely on a sufficient condition: a non-cooperative game has a unique equilibrium if the nonnegative weighted sum of the payoff function is diagonally strictly concave (79).

Definition 2. (Diagonally Strictly Concave) A weighted sum function $h(\mathbf{x}, \mathbf{r}) := \sum_{i=1}^{n} r_i \zeta(\mathbf{x})$ is called diagonally strictly concave for all vector $\mathbf{x} \in \mathbb{R}^{n \times 1}$ and fixed vector $\mathbf{r} \in \mathbb{R}^{n \times 1}$, if for any two different vectors $\mathbf{x^0}$, $\mathbf{x^1}$, we have

$$\Lambda(\mathbf{x^0}, \mathbf{x^1}, \mathbf{r}) \triangleq (\mathbf{x^1} - \mathbf{x^0})^T \sigma(\mathbf{x^0}, \mathbf{r}) + (\mathbf{x^0} - \mathbf{x^1})^T \sigma(\mathbf{x^1}, \mathbf{r}) > 0$$

where $\sigma(\mathbf{x}, \mathbf{r})$ is called pseudo-gradient of $f(\mathbf{x}, \mathbf{r})$, defined as:

$$\sigma(\mathbf{x}, \mathbf{r}) \triangleq \begin{bmatrix} r_1 \frac{\partial \zeta_1}{\partial x_1} \\ \vdots \\ r_n \frac{\partial \zeta_n}{\partial x_n} \end{bmatrix} \tag{4.21}$$

Definition 1 states that the BNE is obtained when all players obtain their best strategies by giving the strategies of other SSs such that their expected payoffs are maximized. Additionally, as stated in Definition 2, the payoff is calculated as the expected value of utility function with the corresponding event probability. In this paper, the event probability is considered as the scheduling class of the connection running on the SS. According to IEEE 802.16 standard, the scheduling class has to be determine during admission control procedure and it cannot be changed after creation of the connection. Consequently, this makes each SS only have one event with probability of 1. Therefore, the description in Definition 1 and 2 match our utility function shown in equation (4.9).

Lemma 2: The weighted nonnegative sum of average payoff u_i in the proposed game is diagonally strictly concave.

proof: We present the weighted nonnegative sum of the average payoff as:

$$h^n(\mathbf{b}, \mathbf{r}) \triangleq \sum_{i=1}^{n} r_i u_i(b_i, b_{-i}) \tag{4.22}$$

where $\mathbf{b} = [b_i \dots b_n]^T$ is the vector of requested bandwidth and $\mathbf{r} = [r_1, \dots r_n]$ is a nonnegative vector assigning weight $r_1, \dots, r_n$ to the average payoffs $u_1, \dots, u_n$, respectively. Similar to equation (4.21), let $\sigma^n(\mathbf{b}, \mathbf{r}) \triangleq [r_1 \frac{\partial u_1}{\partial b_1}, \dots, r_n \frac{\partial u_n}{\partial b_n}]^T$ be the pseudo-gradient of $h^n((\mathbf{b}, \mathbf{r}))$. Each SS $i \in N$ serves a connection belonging to one of scheduling classes. Suppose b_i^{rt}, b_i^{nrt} and b_i^{be} are the amount of bandwidth for the connection in rtPS, nrtPS and BE, respectively. Note that each connection only has one scheduling class. Thus, only one of $b_i^t, t \in \{rt, nrt, be\}$ can be a positive number and the rest of them must be zero. Suppose SS_i serves a rtPS connection, then the average payoff u_i can be actually transformed into

a weighted sum function as follows

$$\begin{aligned} u_i(b_i, b_{-i}) &= \sum_{\gamma} w_\gamma (SI(b_i) - PI(b_i, b_{-i})) \\ &= \sum_{\gamma} w_\gamma \Bigg[\frac{1}{1 + \exp(-k_i^{rt}(b_i - m_i))} \\ &- \Big(TBD(b_i, b_{-i}) \Big)^{\beta_i^{rt}} \cdot \left(\frac{b_i \cdot t_i^{max}}{Q_i^{in} + Q_i^{f}} \right)^{\varphi_i^{rt}} \Bigg] \end{aligned} \tag{4.23}$$

where γ represents the index for different jointly probability events with corresponding probability w_γ. Similarly, the average payoff function can be presented as following if the connection belongs to nrtPS or BE.

$$\begin{aligned} u_i(b^s, b_-^s) &= \sum_{\gamma} w_\gamma (SI(b^s) - PI(b_i, b_-^s)) \\ &= \sum_{\gamma} w_\gamma \Big[k^s \log(1 + h^s b^s) \\ &- \Big(F^s \Big)^{\beta^s} \cdot \Big(U(b^s)^{-1} \Big)^{\varphi^s} \Big] \end{aligned} \tag{4.24}$$

where $s \in \{nrt, be\}$. Now, we can write the pseudo-gradient σ^n as:

$$\sigma^n(\mathbf{b}, \mathbf{r}) = \begin{bmatrix} r_1 \frac{\partial u_1}{\partial b_1} \\ \vdots \\ r_n \frac{\partial u_n}{\partial b_n} \end{bmatrix} \tag{4.25}$$

To check the diagonally strictly concave, we let $\mathbf{b^0}, \mathbf{b^1}$ be two different vectors and define

$$\Lambda(\mathbf{b^0}, \mathbf{b^1}, \mathbf{r}) \triangleq (\mathbf{b^1} - \mathbf{b^0})^T \sigma^n(\mathbf{b^0}, \mathbf{r}) + (\mathbf{b^0} - \mathbf{b^1})^T \sigma^n(\mathbf{b^1}, \mathbf{r}) \tag{4.26}$$

Suppose b_i^0, b_i^1 and r_i are the elements in $\mathbf{b^0}$, $\mathbf{b^1}$ and $\mathbf{r}$ for SS_i, respectively. We want to show that $\lambda(b_i^0, b_i^1, r_i) \in \Lambda > 0$.

$$\begin{aligned} \lambda(b_i^0, b_i^1, r_i) &= (b_i^1 - b_i^0)\sigma(b_i^0, r_i) + (b_i^0 - b_i^1)\sigma(b_i^1, r_i) \\ &= (b_i^1 - b_i^0)[\sigma(b_i^0, r_i) - \sigma(b_i^1, r_i)] \\ &= (b_i^1 - b_i^0)[(\frac{\partial SI(b_i^0)}{\partial b_i^0} - \frac{\partial SI(b_i^1)}{\partial b_i^0}) \\ &+ (\frac{\partial PI(b_i^1)}{\partial b_i^1} - \frac{\partial PI(b_i^0)}{\partial b_i^1})] \end{aligned} \tag{4.27}$$

In Lemma 1, we have shown that the SI is concave for both nrtPS and BE connection. Moreover, for rtPS connection, we also proved that the SS never request bandwidth less than the MSR of the connection. Therefore, only concave part of SI needs to be considered. Without loss of generality, we assume $b_i^1 > b_i^0$. It can lead to $(\frac{\partial SI(b_i^0)}{\partial b_i^0} - \frac{\partial SI(b_i^1)}{\partial b_i^0}) \geq 0$. PI for all scheduling classes is modeled by the exponential function which is strictly convex to the amount of reserved bandwidth. Thus, we can have $(\frac{\partial PI(b_i^1)}{\partial b_i^1} - \frac{\partial PI(b_i^0)}{\partial b_i^1}) > 0$. It results $\lambda(b_i^0, b_i^1, r_i) > 0$. Consequently, we can conclud:

$$\Lambda(\mathbf{b^0}, \mathbf{b^1}, \mathbf{r}) = \sum_{\forall i} \lambda(b_i^0, b_i^1, r_i) > 0 \tag{4.28}$$

It shows that the weighted nonnegative sum of average payoff is diagonally strictly concave. The proof for uniqueness of BNE is completed.

4.8.3 A Note on Framework Implementation

The centralized scheme is assumed that the BS has the completed traffic and queuing information from each SS for optimal bandwidth allocation. This may introduce a large amount of message exchange between BS and SS such that the BS can gather the latest information from the SS in real time.

One of our motivations for studying BR game framework is the amicability for distributed implementation. Unlike the centralized scheme, in our proposed distributed scheme, each SS performs its own operation independently and does not require to gather information from other SSs. Therefore, this proposed game does not introduce any additional message exchanges. The decision made by each SS for bandwidth request is based (as mentioned in Section 6.2) on its belief which is a "guess" on the the strategies selected by other SSs (i.e., the scheduling class and the amount of requested bandwidth of other SSs). This belief does not require the exact traffic information from other SS. The SS determines the amount of requested bandwidth as be best response to its belief.

In IEEE 802.16 network, each SS receives UL-MAP in every MAC frame. It contains the information of bandwidth allocation for each SS. In the proposed game, each SS records

the history of bandwidth allocation for other SSs by extracting it from UL-MAP. The SS constructs its belief based on this history. Note that UL-MAP is a standard message in IEEE 802.16 networks and does not consider as additional message exchanges.

To implement the description above, we design a distributed algorithm shown in Algorithm 7. Each SS i captures the information of other SS j independently, where $j \in N, j \neq i$, and estimates the bandwidth demand of eacg SS j by calculating the expected value of past bandwidth allocation history of SS j from the UL-MAP. The total bandwidth demand without considering the bandwidth requested by i can be estimated by the summation of bandwidth demand of each SS j. The SS i can determine the amount of bandwidth to be requested based on this summation, its QoS requirements and the corresponding utility function defined in Section 4.7.

Algorithm 7 Distributed Algorithm

Input: B_i is the set of possible strategies selected by SS i.
$H_j = \{(p_j^m, h_j^m)\}$ is the set of allocated bandwidth h_j^m with corresponding probability p_j^m for each SS $j \in N$, $j \neq i$.

Output: The amount of requested bandwidth for SS i

1. **For each** j **do**,
   Calculate $b_j = \sum_{\forall m} p_j^m \cdot h_j^m$
2. Calculate $TBD_i = \sum_{\forall j} b_j$
3. Adopt an utility function corresponding to the scheduling class of i.
4. $\forall b_i \in B_i$,
   Find a b_i such that payoff is maximized.

Application	A1	A2	A3
Scheduling Class	rtPS	nrtPS	BE
Minimum Traffic rate (bps)	2.05M	512 k	0
Maximum Sustained Rate (bps)	3.3M	25M	30K
Maximum delay (Sec.)	0.5	1	1
A1: Video Streaming A2: FTP A3: Web Browsing			

Table 4.1 Traffic Parameters

4.9 Numerical and Simulation Results

The numerical analysis is used to evaluate the centralized scheme and the BR game. We adapt the numerical results for the centralized scheme as the optimal solution and compare this to the numerical results for the BR game. The price of anarchy for the BR game is defined as the difference between two schemes. In this section, we first introduce the system model for our numerical analysis and then present our numerical results.

4.9.1 System Model

We evaluate the amount of bandwidth requested for connections in each type of scheduling classes (i.e., rtPS, nrtPS and BE) in our evaluation. The traffic for each connection is assumed to be saturated and the size of queue for each connection is corresponding to the minimum traffic rate and maximum delay requirement. We assume that the bandwidth capacity supported by the BS is 120 Mbps. We evaluate the network with different number of SSs, illustrating different TBDs. Each SS serves one connection at a time. The connection is mapped into either rtPS, nrtPS or BE with equal probability. The start and end time of a connection is randomly generated. We also assume that each SS transmits data via the same modulation. The characteristics of traffics in each scheduling class are summarized in Table 4.1.

Our numerical analysis and simulation are implemented by JAVA as well as AMPL (82), an optimization solver. We consider realistic environments to include various traffic models as shown in Table 4.1 and the request-grant scheme. We evaluate 2000 frames and create

a data generator to simulate data arrival of each connection from the upper layer for both centralized scheme and BR game. It leads to the same amount of arrival data for each connection in every frame for both centralized scheme and BR game. In the centralized scheme, we implement the linear programming formulation shown in Section 4.5 as the input of AMPL. We solve the linear program in each frame and obtain the optimal solution for each connection as well as the overall system utilization.

On the other hand, in BR game, we implement the utility function for each scheduling class specified in Section 4.7. In each frame, the SS tries to determine the bandwidth reservation such that the payoff is maximized. The BS plays a role to enforce the penalty. It calculates the metric for each connection based on the bandwidth utilization of connection and TBD. Then, the BS allocates less bandwidth to the connection with high PI. Again, we simulate a IEEE 802.16 network for 2000 frames. In each frame, the amount of data arrived from the upper layer is same as the data for centralized scheme. We compare the overall system utilization as well as the average throughput for individual connection in both centralized scheme and BR game. The detail of comparison is presented in the following subsection.

4.9.2 Numerical Analysis

In our analysis, we evaluate both centralized scheme and BR game under different TBDs. Fig. 4.2 presents the comparison in overall bandwidth utilization. The bandwidth utilization is presented as the ratio of the summation of allocated bandwidth used by each SS to the total bandwidth that the BS can support. Due to the optimality of centralized scheme, we use "Opt" to represent the numerical results of centralized scheme. Moreover, we evaluate BR game with 4 sets of PI parameters (i.e. β and φ). We mark the results for each set of PI parameter as "p-q-r", where p, q and r stand for the values of PI parameters for rtPS, nrtPS and BE connection, respectively.

In Fig. 4.2, We can observe the difference of bandwidth utilization between these two

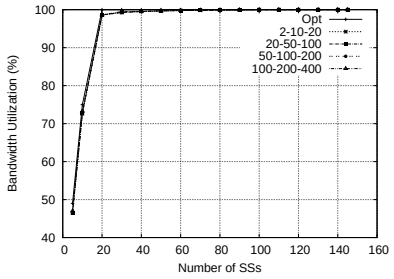


Figure 4.2 Comparison of Bandwidth Utilization

schemes is very limited. Additionally, the SS is able to utilize more bandwidth adapted to the total bandwidth demand in the network. Consequently, the bandwidth is almost fully utilized for the cases after 20 SSs in the network. Moreover, surprisedly, the BR game with different parameters for PI achieve similar bandwidth utilization. It is because each SS tries to maintain a good bandwidth utilization record such that the payoff is maximized. It makes PSI close to 1 no matter what value of φ is used. The amount of requested bandwidth is determined by BDI which is based on TBD. When the network has a low TBD, the SS tries to request bandwidth as much as possible to minimize the queue length. When TBD gets high, the SS would request less bandwidth to reduce the value of PI. It results the TBD close to 1 and leads to similar BDIs when different values of β are used.

To have a better presentation for the bandwidth utilization, Fig. 4.3 shows the difference between the proposed game and the optimal solution. This difference is known as the price of anarchy. The price of anarchy is defined as the performance loss due to the lack of central authority. Although the SS in BR game emphasis on maximizing the payoff, the SSs requires to maximize its bandwidth utilization in order to achieve the maximum payoff. The overall bandwidth utilization is maximized when all SSs reach their maximum payoff. Thus, both centralized scheme and BR game have the same objective of maximizing bandwidth utilization. We present the price of anarchy as the percentage of the optimal solution, which is formally defined as:

$$\frac{BU\_Game - BU\_Optimal}{BU\_Optimal} \times 100\%$$

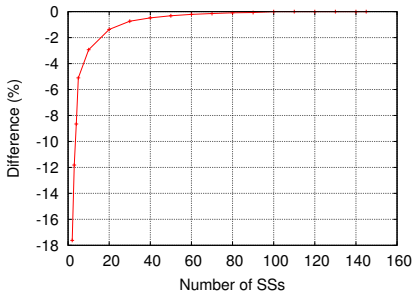
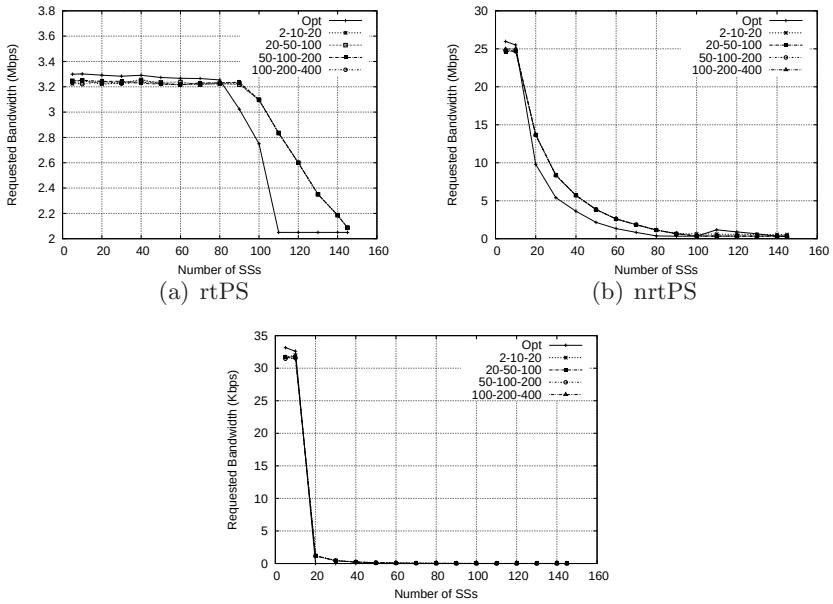


Figure 4.3 Price of Anarchy

where BU_Game and $BU\_Optimal$ stand for the bandwidth utilization achieved by the proposed game and the centralized scheme, respectively. Furthermore, as mentioned earlier that the difference parameters for PI achieve similar bandwidth utilization. Thus, the difference shown in Fig. 4.3 is based on the results from the proposed game with PI parameters "50-100-200". From the figure, we can observe that the price of anarchy achieves around 18% when the number of SS is extremely small. At this time, the amount of requested bandwidth really depends on the queued data. It makes TBD hard to be predicted and the centralized scheme is recommended at this time. The price of anarchy is very limited when the number of SS gets large. We can conclude that in general,the proposed game can almost achieve as good bandwidth utilization as the centralized scheme does.

As mentioned earlier, each connection focuses on maintaining the highest bandwidth utilization (i.e., PSI close to 1). Therefore, the amount of bandwidth requested for the connection is determined by the TBD in the network. Fig. 4.4(a), 4.4(b) and 4.4(c) show the average bandwidth request for rtPS, nrtPS and BE connections, respectively. We can observe that the amount of requested bandwidth for the connection in each scheduling class declines with the increase of TBD. In particular, as shown in Fig. 4.4(a), the amount of requested bandwidth for rtPS connections is relatively stable among all scheduling classes. This matches our discussion in Section 4.7.4 that rtPS connections require stable bandwidth reservation to satisfy the delay requirement. On the other hand, nrtPS connections request the highest amount of bandwidth when the network has a light load. This amount drops



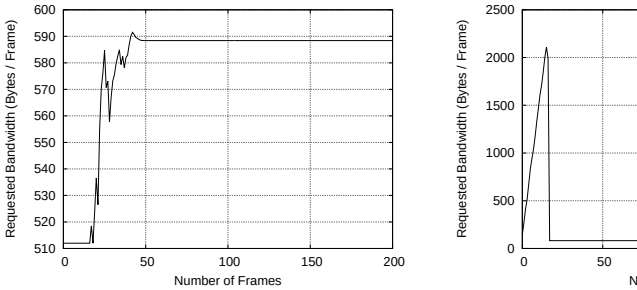
(a) rtPS (b) nrtPS

(c) BE

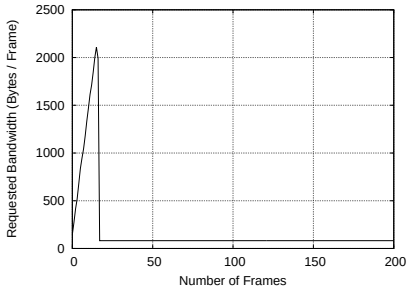
Figure 4.4 Throughput Evaluation

sharply with the increase of TBD. It is worth noting that the bandwidth request for both rtPS and nrtPS connections becomes constant when the number of SS is large enough. It is because our admission control policy ensures the MSR of each connection. Therefore, rtPS and nrtPS connections should request at least their MSR with the increase of number of SSs. The numerical results for BE connection are shown in Fig. 4.4(c). We can observe that the amount of bandwidth requested by a BE connection is nearly zero when the network gets heavily loaded due to flexible QoS requirements.

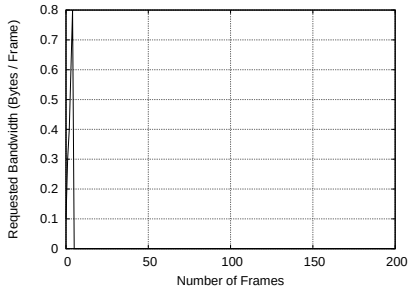
We also investigate how fast the SS determines the stable bandwidth reservation in the BR game. This is considered as the cost for the distributed approach to reach stable status. We analyze the network with 130 SSs and evaluate 1000 frames. Since the bandwidth reservation does not change once stabilized, we focus on the first 200 frames to see how fast it becomes stabilized. Again, the value of parameters for PI (i.e., β and φ) does not affect the amount of requested bandwidth significantly. The results shown in Fig. 4.5(a), 4.5(b) and 4.5(c) are the bandwidth reservation for rtPS, nrtPS and BE connection in the



(a) rtPS



(b) nrtPS



(c) BE

Figure 4.5 Converge of bandwidth reservation for connections in each scheduling class

proposed game with PI parameters "50-100-200", respectively. As shown in these figures, the bandwidth reservation is not stable in the first 50 frames. It is because the SS has not collected enough traffic information for learning the accurate bandwidth reservation made by other SSs. Therefore, it is difficult for the SS to determine the optimal bandwidth reservation for its connection at this stage. However, the bandwidth reservation becomes stable after the SS collects enough traffic information.

4.10 Conclusion

Bandwidth reservation allows quality of service (QoS) guaranteed services to be provided in IEEE 802.16 networks. The BS performs bandwidth allocation based on the amount of bandwidth requested by the SS. Therefore, it is a critical issue to help the SS determine the optimal bandwidth reservation. The objective of this paper is to maximize the system bandwidth utilization while providing QoS guaranteed services. We first construct a centralized scheme by linear programming. In this scheme, we assume that the BS has the completed

traffic information of each SS to perform the optimal bandwidth allocation such that the system bandwidth utilization is maximized with QoS guaranteed services. However, this scheme may require additional message exchanges to let the BS always have the completed information.

We further design a game theoretic approach to help SS make bandwidth reservation without introducing additional message exchanges. We consider both QoS requirements of each connection and total bandwidth demand (TBD) in the network. The objective is achieved when each SS serves its connection with the best achievable service while satisfying QoS requirements. In our game formulation, we model this problem as a distributed bandwidth reservation (BR) game. This game is classified as an non-cooperative Bayesian games. The utility function of the game comprises two indexes: satisfaction index (SI) and penalty index (PI), representing QoS satisfaction and cost for bandwidth reservation of the connection, respectively. Due to different QoS requirements and traffic characteristics, the utility function is customized for each scheduling class. We also investigate the existence and uniqueness of Bayesian Nash Equilibrium (BNE).

In our numerical analysis, we investigate the centralized scheme as well as the BR game with different TBDs in terms of system bandwidth utilization, the amount of bandwidth requested for the connection in each scheduling class as well as the price of anarchy of BR game. The numerical results show both centralized scheme and BR game can reach the similar bandwidth utilization. Additionally, we also confirm that the SS requests more bandwidth for its connection when the TBD is low. With the increase of TBD, the bandwidth for rtPS connections stays relatively stable to satisfy the QoS requirement but the one for both nrtPS and BE connections decreases significantly due to flexible QoS requirements.

CHAPTER 5. Economical Data Transmission in Dynamical Fractional Frequency Reuse

A paper to be submitted IEEE Transactions on Vehicular Technology

David Chuck and J. Morris Chang

Abstract

Dynamically fractional frequency reuse (DFFR) allows the base station (BS) to not only utilize all available frequency partitions but also dynamically adjust the transmission power for each frequency partition corresponding to the current needs. Due to the inter-cell interference, the power allocation in each cell critically affects the throughput in other cells. Thus, how to perform power allocation for each frequency partition becomes a critical issue. This allocation is performed based on not only the traffic demand in the cell but also the power allocation in other cells. In this paper, we focus on the power allocation problem in DFFR and propose an objective of achieving the most economical way for data transmission. To reach the objective, we design a performance objective of maximizing the system throughput per power unit. We believe that this objective matches the desired needs of wireless carriers. We first formulate this problem as an integer linear programming (ILP) problem for optimal solution. Due to high computation complexity of ILP, a greedy algorithm is further proposed as a practical solution. We implement both schemes in our simulation via CPLEX and JAVA program. Our simulation results show that the greedy algorithm has less than 0.2% difference comparing to the results obtained from ILP. We further implement two conventional objectives in our simulation and compare the simulation results with the proposed objective.

5.1 Introduction

With the popularity of mobile multimedia streaming applications such as NetFilx, the Internet traffic has grown dramatically. In addition to qualitative increase, this type of traffic usually has strict quality of service (QoS) requirements. It makes bandwidth demand on the Internet too heavy to be served by the existing network facilities. Consequently, wireless carriers (e.g., AT&T and Verizon) are seeking solutions to enhance system throughput to support the growing traffic demand. In addition to enhance system throughput, recently wireless carriers start to focus on reducing power consumption. This reduction benefits not only wireless carriers to lower the operation cost but also our environment. Therefore, this motivates wireless carriers to purse the most economical method for data transmission which includes the features of enhanced system throughput and limited power consumption. However, generally speaking, achieving high system throughput usually results in large power consumption. Consequently, there is a trade-off between enhancing system throughput and power conservation. We are motivated to investigate this trad-off to achieve the most economical data transmission. Moreover, it is worth noting that providing quality of service (QoS) guaranteed services has become a fundamental requirement in the next generation network. Consequently, we targets the issue of balancing the trad-off between these two desired goals (i.e., enhancing system throughput and reducing power consumption) while satisfying QoS requirements of all applications.

The forth generation (4G) networks might become a possible solution for wireless carriers to enhance system throughput. It aims to support high bandwidth, large coverage and QoS guaranteed services. Currently, WiMAX (1) (53) and LTE (83) are two major wireless technologies in 4G networks. In the development of 4G networks and advanced version (e.g., IEEE 802.16m), there are two directions specified in these technologies to enhance the system throughput: wireless medium access technologies and spectrum efficiency. In the first direction, several advanced wireless medium access technologies such as multiple-input and multiple-output (MIMO) and Orthogonal Frequency-Division Multiple Access

(OFDMA) are adopted in 4G networks. These technologies help 4G networks support high transmission rates, wide service coverage and large bandwidth capacity. Although the benefits brought by these technologies enhance the system throughput, it is required to improve the spectrum efficiency in order to achieve the maximum system throughput. The spectrum efficiency describes the number of cells that a frequency partition is used. Larger spectrum efficiency represents that more spectrum available in each cell. Therefore, higher system throughput is achieved when the spectrum efficiency is larger.

Fractional frequency reuse (FFR) (84) has been introduced in 4G technologies to enhance the spectrum efficiency. It allows all available frequency partitions to be utilized in each cell with different transmission power. Due to unequal transmission power for each frequency partition, FFR is able to improve the spectrum efficiency without experiencing significant inter-cell interference. However, the coverage of each frequency partition in FFR is preplanned and cannot be customized to the current traffic demand and user distribution which change dynamically. Therefore, this inflexibility may lead to suboptimal system throughput and also cause the wastage of transmission power due to maintaining the fixed coverage larger than the current needs. Therefore, an improvement is needed to maximize the system throughput and reduce the power consumption.

Fortunately, an advanced version, named dynamical FFR (DFFR) (85), has been proposed to allow the transmission power of each frequency partition to be adjusted dynamically based on the current needs. Although this flexibility allows the base station (BS) to operate with customized power allocation, how to allocate appropriate transmission power to each frequency partition becomes a critical issue in DFFR. As stated earlier, all frequency partitions are utilized in each cell. Thus, the decision of power allocation for each frequency partition made in each cell affects not only the system throughput in that cell but also in other cells due to inter-cell interference. Consequently, a comprehensive decision is needed for each BS while performing power allocation. This decision should be based on not only the current needs of the cell but also the power allocation of other cells. To

gather this information, the BS may need to exchange the information of power allocation with each other in order to make an appropriate decision. However, this may require a large amount of message exchange between BSs, which is considered as the overhead in a network. Consequently, a clever scheme of power allocation is desired to help the BS allocate the transmission power to each frequency partition dynamically with limited network overhead.

There have been several research works regarding performance analysis of DFFR (85)-(95). Among these works, two popular performance objectives can be concluded: 1) maximizing network throughput. 2) minimizing power consumption. The first objective does not consider the cost of power consumption while pursuing maximum network throughput. Generally, achieving higher throughput usually consumes more transmission power. However, the relation between throughput and power consumption may not be linear due to varied channel quality and inter-cell interference. It may lead to consume a large amount of power for limited throughput improvement. On the other hand, the second performance objective aims to minimize power consumption in each frame without considering the degradation of system throughput. Although the consumed power is minimized in each frame, it may result in longer transmission time to transmit the same amount of data. Thus, the total energy consumption to complete the task may not be minimized. In summary, these performance objectives only deal with either one of goals that wireless carriers are pursuing. Consequently, with the current performance objectives, it is still difficult for DFFR to achieve these goals at the same time.

In this paper, we focus on the power allocation problem in DFFR and aim to balance the trade-off between system throughout and power consumption with the consideration of QoS guaranteed services. We proposed a metric, named efficiency ratio, which is the ratio of the system throughput to the cost of power consumption. This ratio represents the system throughput contributed by per cost unit of power consumption. Larger efficiency ratio indicates more throughput contributed by per cost unit. The objective of this paper is

to allocate the transmission power to each frequency partition such that the efficiency ratio is maximized. In addition, we also ensure the QoS requirement of each SS. In summary, we aim to perform transmission power allocation to maximize the efficiency ratio with QoS guaranteed services. To the best of our knowledge, we are the first to propose this objective for the performance analysis of DFFR.

As mentioned earlier, the BS may rely on a large amount of message exchange to gather the actual information of power allocation in other cells to make its decision. Although the information may be obtained through the channel quality feedback by each subscriber station (SS) to avoid the network overhead, the information may not be accurate enough if more than one BS performs power allocation at the same time. To alleviate this issue, we adopt a backoff mechanism to improve the probability that only one BS adjusts the transmission power in each frame. With the help of this backoff mechanism, each BS can perform its power allocation in a distributed fashion based on the channel quality estimated by the SS and focus on maximizing the efficiency ratio with QoS guaranteed service.

As our definition, the efficiency ratio is the ratio of system performance to the cost of power consumption. To maximize this, we transfer the objective as maximizing the system throughput while minimizing the power cost. Each combination of transmission power allocated to the frequency partition results in the corresponding system throughput and power cost, respectively. We formulate this objective with a "benefit-cost" concept. The benefit is the system throughput that the cell receives and on the other hand, the corresponding transmission power cost is the cost in the concept. The payoff is defined as the difference between the benefit and cost (i.e., benefit $-$ cost). We first formulate this power allocation problem by integer linear program (ILP). This formulation leads us to the optimal power allocation such that the payoff is maximized. Due to high computation complexity, the ILP turns out to be impractical over any reasonably large case.

We further propose a heuristic algorithm based on greedy approach as a practical solution. We compare the results from both ILP and heuristic algorithm through simulation. In

our simulation, we generate several test cases in terms of the number of SS per cell as well as the number of cell in the system. For each test case, we implement the ILP and heuristic algorithm through CPLEX 10.2 and JAVA programming. The simulation results show that the heuristic algorithm can reach less than 0.2% difference comparing to the optimal solution from ILP. We further implement two conventional performance objectives: minimizing power consumption (MIN-Power) and maximizing system throughput (MAX-Throughput). Our simulation results show that the proposed scheme can achieve the highest efficiency ratio which is the ratio of system throughput to power consumption cost.

In summary, the contribution of this paper can be listed as follows. First, we propose a new performance objective for the problem of power allocation in DFFR. This performance objective matches the feature of an ideal network currently desired by wireless carriers. Moreover, we formulate our power allocation problem as an integer linear programming problem and solve it by CPLEX to obtain the optimal power allocation. Further, due to high computational complexity, we propose a heuristic algorithm based on greedy approach as a practical solution. We implement both ILP and heuristic algorithm through simulation and compare their simulation results. We further compare the simulation results of the proposed objective with two conventional performance objectives.

The rest of this paper is organized as follows. We present a brief introduction of frequency reuse as the background information in Section 5.2. In Section 5.3, we present the related works proposed in the literature. The system model used in this paper is presented in Section 5.4. The ILP formulation followed by our greedy algorithm is shown in Section 5.5 and Section 5.6, respectively. We present our simulation results of both ILP and greedy algorithm in Section 5.7. Finally, the conclusion is located in Section 5.8.

5.2 Background information

In addition to advanced wireless medium access technologies, frequency assignment is another key factor to boost up the network throughput successfully. The frequency parti-

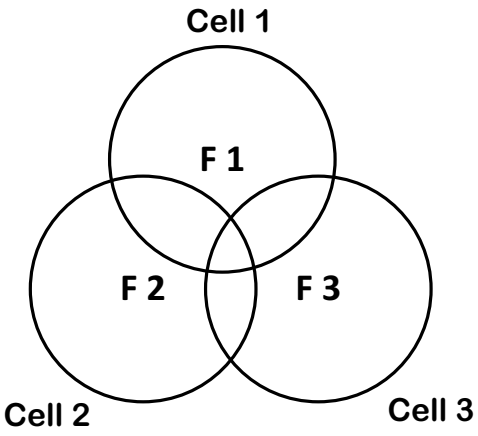


Figure 5.1 Frequency Reuse

tion assigned to each cell should be planned in order to avoid throughput degradation due to inter-cell interference. The conventional frequency assignment mechanism, named frequency reuse, divides the available frequencies into several disjointed frequency partitions. Each frequency partition is assigned to one cell and the adjacent cells must utilize different frequency partitions to avoid inter-cell interference. An example is shown in Fig. 5.1. In the figure, Cell 1, 2 and 3 use frequency section F1, F2 and F3, respectively. Since there are no overlapped frequency partitions between cells, this mechanism successfully limits inter-cell interference. However, due to the disjointed frequency partition allocation, each cell can only utilize one third of all available frequencies (i.e, one frequency partition). This leads to low spectrum efficiency. For example, the spectrum efficiency in Fig. 5.1 is 1 since each frequency partition is only available in one cell.

Recently, a scheme has been introduced to improve the spectrum efficiency by allowing all available partitions to be utilized in every cell but operated in different transmission power. Because each cell utilizes a fraction of available partitions, this scheme is called fractional frequency reuse (FFR). Since all frequency partitions are available in each SS, the BS has wider spectrum to serve its users with high bandwidth capacity. Moreover, the allocation of unequal transmission power helps FFR alleviate inter-cell interference. Conse-

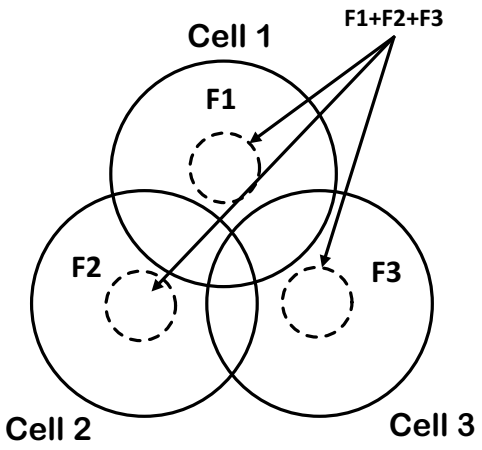


Figure 5.2 Frequency Reuse

quently, FFR is able to improve the spectrum efficiency with limited inter-cell interference. For example in Fig. 5.2, each cell utilizes two frequency profiles: solid and dashed circles. The solid circle represents the coverage of a frequency partition assigned to the cell. The dashed circle, on the other hand, stands for the coverage of utilizing all available frequency partitions (i.e., F1+F2+F3). Since each cell has adjusted the transmission power of all frequency partitions properly, there is no inter-cell interference between cells and the spectrum efficiency is improved to 3.

Although FFR improves the spectrum efficiency, the transmission power allocated to each frequency partition is preplanned and cannot be adjusted dynamically. However, the traffic demand and user distribution in each cell may change from time to time. This inflexibility may prevent FFR to customize the service coverage to fit the current needs of the cell. For example, suppose all SSs in Cell 1 shown in Fig. 5.2 locate within the dashed circle. The BS in Cell 1 can shrink the coverage of F1 to lower power consumption. However, due to preplanned power allocation, the coverage of F1 is fixed. This fixed coverage not only causes the wastage of transmission power but also prevent other cells (i.e., Cell 2 and 3) to enlarge their coverage to achieve higher bandwidth capacity. Therefore, preplanned power allocation in FFR may not be able to maximize the system throughput when each

cell has unequal traffic demand.

Recently, dynamic fractional frequency reuse (DFFR) has been proposed to allow the BS to adjust transmission power dynamically. Due to inter-cell interference, this adjustment is performed based on not only the traffic demand but also the power allocation in other cells. The BS in each cell can work together and customize its power allocation for the current need. Following the example above for FFR, the BS in Cell 1 can shrink it coverage of F1 to conserve its transmission power. At the same time, other cells (i.e., Cell 2 and 3) experience less interference on F1 due to weak transmission power of Cell 1. Thus, they can enlarge its coverage of F1 to achieve higher system throughput. In this paper, we focus on power allocation problem in DFFR. We aim to determine the transmission power dynamically for each frequency partition to achieve our objective of maximizing the efficiency ratio with QoS guaranteed services.

5.3 Related Works

Among the existing research works, two performance objectives of performance analysis in FFR can be concluded: 1) maximizing network throughput. 2) minimizing transmission power. In (89), Stolyar *et. al.* proposed a distributed mechanism for power allocation in DFFR. They focus on Best Effort (BE) traffic and aim to maximize the network utility which is a function of average transmission rate. Since only BE traffic is considered, this work does not consider other types of traffic with QoS requirements such as multimedia streaming. In (90), Ali-Yahiya *et. al.* investigate an architecture that coordinates the allocation of resource. They consider the trade off between maximizing system and QoS requirement. However, the cost of power consumption is not on their list. Thus, it may end up spending a large amount of power for very limited throughput improvement. On the other hand, the second performance objective aims to minimize the power consumption in each frame without considering the degradation of network throughput. It may result in more transmission time and total energy consumption to transmit the same amount of

data. In (92), the authors investigated a dynamic fractional frequency reused proportional fair in time and frequency scheduling which considers fairness in time, frequency and user dimensions. Both centralized and decentralized scheme were proposed to improve long term system throughput.

The second performance objective focuses on minimizing transmission power. In (100), a scheme for joint allocation of modulation scheme, coding rates, resource blocks and power has been proposed for LTE networks. This work tries to find a combination such that the transmission power is minimized. Moreover, the QoS requirement of different types of traffic is also considered in this work. However, since this works aims to minimize the power consumption in each frame, the authors only satisfy the minimum QoS requirement of traffic. This may take longer time to finish the same task and consume more transmission power. Another work to minimize the power consumption is proposed in (88). In this work, the authors proposed an approach based on continuous "selfish" optimization of resource allocation by each sector. Their analysis and simulation show that the proposed algorithm leads the system to "self-organize" into efficient frequency reuse pattern. However, the authors only focus on constant bit rate (CBR) traffic. Therefore, the traffic with different QoS requirement is not considered by the authors. In (94), the spectral efficiency is evaluated. However the authors only consider the minimum power allocation with QoS guaranteed services. With these conditions, the most economical data transmission may not be achievable.

In additional, instead of using system level simulations using a hexagonal grid for the base station locations, Novlan *et al.* use a Poisson point process to mode the base station locations (96). By using this, they evaluated both static and dynamical FFR. Their results provide insight into system design guidelines. Based on the description above, the existing work focuses on either maximizing system throughput or minimizing power transmission. However, none of them emphasis on the most economical data transmission. Moreover, providing QoS guaranteed service should be also a fundamental feature. In this paper, we

are pursuing the most economical data transmission and focus on an objective of maximizing the efficiency ratio with QoS guaranteed services.

5.4 System Model

Suppose there are $|N|$ cells in our system, where $N = \{1, 2, \ldots, n\}$ represents the set of cells. We assume that for any cell $i \in N$, it comprises one BS and $|S_i|$ SSs, where $S_i = \{s_1^i, s_2^i, \ldots, s_{k_i}^i\}$ is the set of SSs in cell i. The SSs are randomly distributed in each cell. The BS in each cell is assumed to be identical in terms of frequency accessibility, power capacity and computation capability and responsible for determining when and how to perform data transmission for each SS.

We adopt OFDMA as the physical layer in our system model, where the spectrum is divided into several subchannels. Moreover, the time domain in a MAC frame is slotted. Thus, the minimum resource unit for a BS to be allocated to a SS is one subchannel per time slot. We call this minimum resource unit as resource block (RB). As our assumption that the BS in each cell is identical, there should be the same number of RBs in each cell. These RBs are represented by a set of RBs, $B = \{1, 2, \ldots, b\}$. Moreover, we assume that the BS supports discrete power level to each RB. These power levels are represented by a set $P = \{p_1, p_2, \ldots, p_L\}$. It is worth noting that p_L indicates the largest power level in P to be allocated to a RB and cannot be larger than the power capacity of the BS.

The objective of each BS is to schedule RBs with the corresponding transmission power to SSs such that the efficiency ratio is maximized and QoS guaranteed service can be provided. To achieve the maximum efficiency ratio, we incorporate a benefit - cost concept, where benefit and cost refer to system throughput and the corresponding power consumption, respectively. The payoff is defined as the difference between system throughput and the corresponding power cost (i.e., throughput minus power cost). The payoff is contributed by the payoff of each RB. For each allocated RB, the corresponding transmission power must be larger than zero. Otherwise, the RB should not be allocated with any transmission

power.

In our system, we consider three types of traffic: real-time, non-real time and best effort. Each type of traffic has its own QoS requirement. Each SS randomly serves up to M applications. Each application can be mapped to one of these types of traffic. It is worth noting that all applications served by SSs must be admitted by the BS before operation in order to ensure that the BS has enough resource to guarantee their QoS requirement. We represent the set of SSs serving real-time, non-real time and BE traffic by S_r^i, S_{nr}^i and S_{be}^i, respectively, where $S_r^i \cup S_{nr}^i \cup S_{be}^i = S^i$. Due to different location of each SS, we assume that the modulation and coding scheme (MCS) used by each SS is predetermined and considered as an input of our system (100).

As mentioned earlier, in DFFR, the power allocation performed in each cell is based on not only its current traffic demand but also the power allocation of other cells. The BS needs the information of power allocation in other cells while scheduling RBs to its SSs. A traditional method to gather this information relies on message exchange between BSs. However, this may cause huge network overhead. To avoid the overhead, the BS may estimate the power allocation of other cells through the channel condition periodically reported by its SSs. However, due to simultaneous power allocation, this information may not be accurate when more than one cell performs their power allocation at the same time.

In order to alleviate this issue, we implement a random backoff mechanism in our system, which is similar to the miulticast polling mechanism used in IEEE 802.16 networks (97). This mechanism contains two stages: backoff stage and allocation stage. In backoff stage, the BS selects a random number between 0 and maximum backoff number W. This backoff counter is deducted by 1 in each frame and indicates the number of frame that the BS should defer. When the counter reaches zero, the BS enters into allocation stage. In this stage, the BS selects another random number between 0 and 1. This random number determines whether the BS performs power allocation in the current frame. The BS compares it with the predetermined threshold T which is also between 0 and 1. If this random number is

larger than T, the BS performs power allocation. Otherwise, the BS reselects a random number between 0 and 1 and doubles the threshold T until the maximum attempt limit has reached. If the BS cannot performs power allocation within the maximum attempt limit, it resets everything and enters into backoff stage again.

Since this backoff mechanism does not guarantee that the BS is able to perform power allocation on time. This may prevent the BS to providing QoS guaranteed services. Consequently, the maximum time interval is needed. If the time duration since last allocation reaches this maximum time interval, the BS should perform power allocation right away to ensure the QoS guaranteed services. Therefore, the frequency of power allocation depends on the selected backoff value and the maximum time interval. Due to randomized backoff counter, this mechanism can effectively reduce the probability that more than one cell performs power allocation in the same MAC frame.

Furthermore, in our scheme, the power allocation does not change until the next allocation is performed. It is possible that the channel quality is different to the one estimated during previous allocation due to effect of multipath and shadowing. When the channel quality gets bad, the lower MCS may be used to ensure successful transmissions. However, due to lower throughput per RB, the number of allocated RBs for a particular SS may not be enough to cover the QoS requirement. On the other hand, the number of allocated RB is still able to cover the QoS requirement when the channel quality gets improved. Therefore, in order to buffer the moderate change of channel quality, in our allocation, we use the MCS which is κ-th level lower than the target MCS for each SS. This concept is employeed to determine the MCS of each SS, used in both ILP and heuristic algorithm. When κ gets large, the problem gets more complicated to be solved. In our simulation, we set κ =1.

5.5 Integer Linear Programming

Because the backoff mechanism reduces the probability that more than one cell updates the power allocation at the same time, the BS can rely on the information of channel

condition reported by its SS to perform power allocation. As mentioned in Section 5.4, RB is the minimum resource unit that the BS can allocate to its SSs. Consequently, a RB cannot be shared by more than one SS. Due to this integrality, we formulate our power allocation problem by integer linear programming (ILP). The detail of our formulation is presented in this section. We first introduce the objective function used in our formulation. As mentioned earlier, there are several physical restrictions in the system such as power capacity of BS and QoS requirements of each SS. We include these restrictions as constraints in our ILP formulation. To have a clear presentation, we summary all parameters used in our formulation in Table 5.1.

The objective of our problem aims to schedule the optimal RB and power allocation in each frequency partition such that the efficiency ratio is maximized with QoS guaranteed service. We adopt a "benefit-cost" concept to model this objective. For each allocated RB, the benefit is the throughput received by the allocated SS and the cost is the price paid for the corresponding power consumption. The payoff is defined as the difference between benefit and cost (i.e., benefit minus cost). The system payoff is the sum of payoff of each individual RB. Consequently, the objective function for our ILP formulation is presented as below:

$$Maximize:$$

$$\sum_{i\in N}\sum_{s_j^i\in S_i}\sum_{b\in B}\sum_{p_l\in P} Y(s_j^i,b,p_l)\cdot\left(R(s_j^i,b)-P_c(s_j^i,b,p_l)\right) \tag{5.1}$$

In (5.1), $R(s_j^i,b)$ and $P_c(s_j^i,b,p_l)$ stand for network throughput and power consumption cost, respectively, when the SS s_j^i in cell $i\in N$ utilizes the RB b with transmission power p_l, where $(s_j^i,b)\in S_i\times B$ and $p_l\in P$. The total payoff contributed by each SS is the sum of payoff contributed by each RB allocated to the SS. Since all SSs in the cell have equal opportunity to utilize each RB, we introduce a binary decision variable $Y(s_i^j,b,p_l)$ for each pair of SS s_j^i and RB b. If SS s_j^i utilizes RB b with transmission p_l, then $Y(s_i^j,b,p_l)$ is set to 1. Otherwise, $Y(s_i^j,b,p_l)$ is 0. Moreover, each SS operates with its own MCS which may

result to different throughput. Thus, the value of $R(s_j^i, b)$ depends on the MCS and may not be same for all SSs.

In addition to the objective function above, we consider all practical requirements as the constraints in our formulation. In practice, it is impossible that the BS has unlimited power to serve its SSs. Thus, we assume that P_{bs}^i is the power capacity of BS for cell $i \in N$ and the summation of all power allocated to each SS cannot be more than this capacity. We call this requirement power capacity constraint and present it as (5.2):

$$\sum_{s_j^i \in S_i} \sum_{b \in B} \sum_{p_l \in P} p_l \cdot Y(s_j^i, b, p_l) \leq P_{bs}^i \qquad \forall i \in N \tag{5.2}$$

Providing QoS guaranteed services is one of important and fundamental features in 4G networks. We include this feature into our problem while pursuing the most economy way for data transmission. This feature is translated as a constraint named QoS constraint listed below:

$$\sum_{b \in B} \sum_{p_l \in P} R(s_j^i, b) \cdot Y(s_j^i, b, p_l) \geq R_{req}^{s_j^i} \qquad \forall s_j^i \in S_r^i \tag{5.3}$$

where

$$R_{req}^{s_r^j} = \frac{Q_r + Q_r^f}{T_{i,j}^{max}} \tag{5.4}$$

In (5.3), $R_{req}^{s_r^j}$ stands for the QoS requirement of SS s_j^r in terms of bytes, where s_j^r refers to the SS serving real-time traffic. Due to delay sensitivity of real time traffic, the BS has to ensure that the maximum delay requirement of real time traffic is satisfied (86)(87). Consequently, $R_{req}^{s_r^j}$ shown in (5.3) is calculated based on the expected queued data and the maximum delay requirement as shown in (5.4). Since in our scheme, the frequency of power allocation depends on the selected random backoff number, the expected queued data is calculated as the current queued data plus the expected amount of data arriving until the next power allocation.

It is not necessary to have strict delay requirement for non-real time and BE traffic. The BS ensures to serve enough bandwidth to satisfy the minimum bandwidth requirement agreed during admission control. Therefore, the QoS requirement for these two types of traffic is presented as below:

$$\sum_{b \in B} \sum_{p_l \in P} R(s_j^{r'}, b) \cdot Y(s_j^{r'}, b, p_l) \geq R_{req}^{s_j^{r'}} \qquad \forall s_j^{r'} \in S_{nr}^i \cup S_{be}^i \tag{5.5}$$

Similar to (5.3), $R_{req}^{s_j^{r'}}$ stands for the QoS requirement of SS $s_j^{r'}$ in terms of bytes, where where $s_j^{r'}$ refers to the SS serving non-real time and BE traffic.

As stated in Section 5.4, each SS has its own MCS which is an input in our problem. Different MCSs require different SINR thresholds in order to be operated. The BS has to allocate enough power to sustain these SINR thresholds. We represent this SINR requirement for s_j^i at RB b as our third constraint named SINR constraint shown in (5.6):

$$\frac{RSS(s_j^i, p(s_j^i, b))}{I_{s_i^j, b}} \geq SINR_{req}^{s_j^i} \tag{5.6}$$

where

$$p(s_j^i, b) = \sum_{p_l \in P} p_l \cdot Y(s_j^i, b, p_l) \tag{5.7}$$

In (5.6), $SINR_{req}^{s_j^i}$ is the SINR threshold of the MCS that SS s_j^i uses. The left side of (5.6) describes the SINR of s_j^i at RB b. $RSS(s_j^i, p(s_j^i, b))$ is the received signal strength for s_j^i corresponding to the transmission power $p(s_j^i, b)$ at RB b, where $p(s_j^i, b)$ represents the transmission power for s_j^i selected by equation (5.7). $I_{s_j^i, b}$ represents the interference strength for s_j^i at RB b. This includes the interference from other cells as well as the background noise. The background noise is assumed as a constant in our system. However, the interference from other cells may be different for each SS depending the distance between the SS and BSs in other cells.

Although all SSs in the cell have equal opportunity to access all RBs, a RB can only be allocated to at most one SS. Moreover, the allocated SS operates in only one level of

transmission power. This requirement is enforced through our constraint in (5.8) as known as the non-sharable constraint in our ILP formulation

$$\sum_{s_j^i \in S_i} \sum_{p_l \in P} Y(s_j^i, b, p_l) \leq 1 \qquad \forall b \in B, i \in N \tag{5.8}$$

Finally, we present our constraint for the variables in our ILP formulation. There is only one boolean variable in our formulation, $Y(s_j^i, b, p_l)$, indicating whether the RB b is allocated to SS s_j^i with transmission power p_l. The variable constraint in our formulation is presented as following:

$$Y(s_j^i, b, p_l) \in \{0, 1\} \qquad \forall s_j^i \in S_i, \forall i \in N, \forall b \in B \tag{5.9}$$

Although ILP leads us to optimal power allocation such that the sum of payoff contributed by each cell is maximized, it turns out to be intractable over any reasonably large inputs. Therefore in the next section we present a simple and fast heuristic algorithm based on greedy approach to solve the power allocation problem.

5.6 Greedy Algorithm

Due to high computation complexity of ILP, we further propose a heuristic algorithm to perform power allocation efficiently. This proposed algorithm is based on greedy approach. Same as our ILP formulation, the objective of our greedy algorithm aims to maximize the total payoff of the cell. Clearly, the constraints specified in our ILP formulation such as QoS requirement, power capacity, and SINR should be also held in our greedy algorithm. The detail of the proposed algorithm is presented in Algorithm 8.

Our algorithm operates in per RB fashion. It means that the BS allocates one RB with the required transmission power to a SS in each time. Initially, the available transmission power of the BS is equal to its power capacity. In each time that a RB is allocated to a SS, the corresponding power consumption is deduced from the available transmission power.

Algorithm 8 Greedy Algorithm

Input: 1. The location of all SSs and BS.
2. Power capacity of BS.
3. SS MCS.

Output: 1. Power allocation
2. Payoff for the cell

Phase I: Investigation:

For each $s_j^i \in S_i$ **do**
For each $b \in B$ **do**
1. For each $p_l \in P$ **do**
a. Calculate the corresponding payoff for each (s_j^i, b, p_l).
b. Record the smallest p_l which can sustain the required MCS.
End For
2. Record a RB b_j which leads to the largest payoff PF_j.
End For
End For

End Phase I.

Phase II: Allocation:

For $j = 1$ to $|S_i|$ **do**
1. Check whether QoS requirement of s_j^i is satisfied or not.
2. If there is at least one SS with unsatisfied QoS requirement.
Do record one with the largest payoff among these SSs with unsatisfied QoS.
Else
Do record one SS with the largest and non-negative payoff
End For

End Phase II.

The algorithm should terminate when the available transmission power cannot support the requirement of SS. There are two phases in our algorithm: investigation and allocation. In the investigation phase, the BS calculates the payoff among all available RBs for each SS and records one with the maximum payoff. As mentioned earlier, the SS can operate in different MCS which is an input of our problem. Therefore, in this phase, the BS focuses on not only maximizing the payoff of each SS but also make sure that it can support enough transmission power to sustain the required MCS.

With the information gathering in the investigation phase, the BS starts to make decisions of allocating RB to SS in the allocation phase. In order to ensure that the QoS guaranteed service can be provided, all SSs are classified into two categories: required and optional. The first category indicates the SS which QoS requirement has not been satisfied. On the other hand, the second category stands for the SS with satisfied QoS requirement. Due to the characteristic of delay sensitivity, the QoS requirement for real time traffic is based on the maximum delay requirement. We use the same method as ILP to calculate this requirement as shown in (5.4). The requirement for non-real time and best effort traffic is based on the minimum guaranteed bandwidth as agreed during admission control since less strict QoS requirement is needed. If there are SSs fallen into the first category, the BS must allocate RBs to these SSs in order to meet the requirement of providing QoS guaranteed service. At this time, the BS starts to select one SS with the largest payoff among the SSs in the first category and allocate the corresponding RB to this SS. After allocating the RB, the BS marks that RB as unavailable and deduce the required transmission power from available power of BS. If there are no SSs in the category of required, it means that the QoS requirement of all SSs has been reached. At this time, the BS can select one SS with the largest payoff and allocate the corresponding RB to that SS. This allocated RB is marked as unavailable and should not be allocated to any other SS in the future.

After allocating a RB to a SS, the BS repeats these two phases until all RBs are unavailable. In addition to no available RBs, this algorithm terminates when one of the following

conditions are met: 1) the BS does not have enough available transmission power. 2) all SSs have negative payoff. The first condition ensures that the BS has enough available transmission power to sever each selected SS. This matches the power capacity constraint shown in (5.2) in our ILP formulation. As stated in (5.1), our objective is to maximize the payoff of the cell. It is necessary to ensure that all allocated RBs contribute positive payoff. Thus, the algorithm should end when no SSs have positive payoff.

Complexity. Our greedy algorithm comprises two phases. Thus, the complexity of this algorithm can be calculated as the sum of complexity of individual phase. In the investigation phase, each SS takes $O(|B|)$ time to go through all RB. Each RB takes $O(|P|)$ time to find the optimal power level. Thus, the total complexity in this phase is $O(|S_i||B||P|), \forall i \in N$. In the allocation phase, the BS takes $O(|S_i|)$ time to go through all SS to meet the requirements in the phrase. Thus, the total time for a BS to allocate one RB is $O(|S_i||B||P|+|S_i|)$ and there are total $|B|$ RBs. Consequently, total complexity for the greedy algorithm is $O(|R| \cdot (|S_i||B||P| + |S_i|))$.

Correctness. The greedy algorithm leads us to a valid power allocation due to the following constraints maintained by the algorithm - 1) The BS ensures that it has enough available transmission power and is able to support the corresponding MCS before allocating a RB. 2) The SS in required category must be served before allocating RB to the SS in optional category. This ensures that the QoS requirement of each SS can be satisfied. 3) Once a RB is allocated to a SS, it is marked as unavailable. it avoids that one RB is shared by more than one SS. 4) All RBs allocated to a SS contribute positive payoff. Thus, this leads us to maximum payoff for the cell. The correctness of the proposed algorithm is verified through simulation presented in the next section.

5.7 Performance Evaluation

In this section, we implement both ILP formulation as well as heuristic algorithm shown in Section 5.5 and 5.6 in our simulation. We first introduce the system model used in our simulation and then compare the simulation results of two schemes. We also implement the conventional performance objectives in our simulation, maximizing system throughput (MAX-Throughput) and minimizing power consumption (MIN-Power), and compare the simulation results with the proposed objective in terms of efficiency ratio which is defined in the later of this section.

5.7.1 System model

We simulate both schemes (i.e., ILP formulation and heuristic algorithm) in the system of 2 and 3 cells, respectively, and the heuristic algorithm in the system of 7 cells since ILP becomes intractable in the system of 7 cells. Each cell serves 5 different numbers of SSs from 10 to 50. These SSs are randomly distributed in the service coverage of the corresponding BS. The purpose of using different number of SSs is to simulate these two scheme under different traffic load. We implement the heuristic algorithm via Java program and compute our ILP formulation by CPLEX 10.2(98).

Each SS randomly serves 1 to 5 applications. Each of them randomly belongs to one type of traffic shown in Table 5.3. It is worth noting that all applications served by each SS must pass the admission control enforced by the BS before operation. The admission control is used to ensure that the BS has the capability to provide the guaranteed resource to each SS. In this paper, we implement the admission control with two aspects: bandwidth capacity of the BS and admission control policy for each SS. In DFFR, the cell can enlarge its coverage only when its adjacent cells have less bandwidth demand. It is possible that all adjacent cells have same bandwidth demand. In this case, the BS should go back to the traditional FFR to alleviate inter-cell interference. Moreover, it is possible that each SS uses the lowest MCS for data transmission. Consequently, in our scheme, the total bandwidth capacity for

the BS is calculated based on the traditional FFR with the lowest MCS. Furthermore, the SS should determine the QoS requirements such as minimum sustain rate and maximum traffic rate based on the characteristic of traffic and specify these requirements in the admission control request during admission control procedure. The BS can either accept or reject the request based on the current available resource and admission control policy. The admission control policy for each SS implemented in our simulation follows the minimum sustained rate. The BS has to ensure this rate to all requests in order to accept it during admission control.

As stated earlier, a backoff mechanism is employed to reduce the probability that more than one cells have power allocation at the same time. Each BS randomly selects its backoff counter between 0 and the maximum backoff windows, W. The value of W is set to 100 in our simulation. It makes the average number of frame that the BS defers its attempt for power allocation is 50. Further, we set the threshold for performing power allocation, T, as 0.3. It means the initial successful probability to perform power allocation is 70%.

As stated in our system model, the MCS used by each SS is an input of our problem. In our simulation, we implement 15 MCSs and the detail of each MCS is presented in Table 5.4 (99). Each SS is using one type of MCS depending on its location. Due to different location of each SS, the interference experienced by each SS is calculated individually. Based on the frequency used in our simulation, we adopt the path loss model interference is show in equation (5.10) (100).

$$L = 128.1 + 37.6 \cdot \log_{10}(R) \tag{5.10}$$

The received signal strength received by each SS from each BS is calculated by this path loss model. It must be at least the SINR threshold corresponding to its MCS in order to have successful operation.

5.7.2 Simulation Results

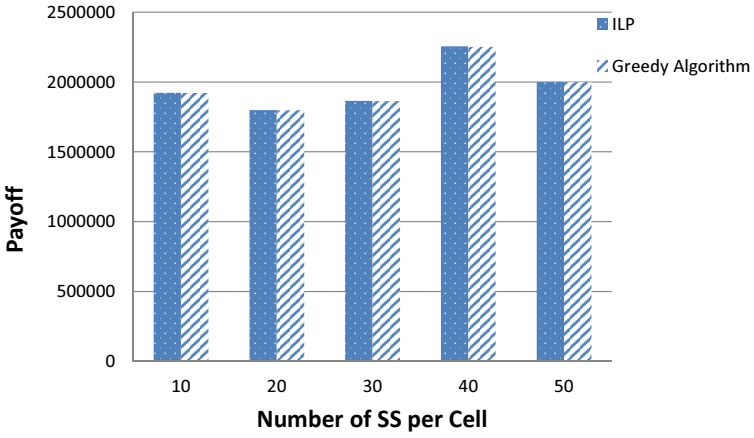
We first present our comparison of simulation results for both ILP formulation and greedy algorithm. This comparison is made in terms of average payoff received for each cell. Fig. 5.3 and 5.4 show the simulation results and comparison for 2-cell and 3-cell environments, respectively. Fig. 5.3(a) and 5.4(a) present the comparison of simulation results between our ILP formulation and greedy algorithm. From the figures, we can observe that the gap between these two schemes is very limited.

We further investigate this gap in terms of the percentage of ILP simulation results, which is calculated as

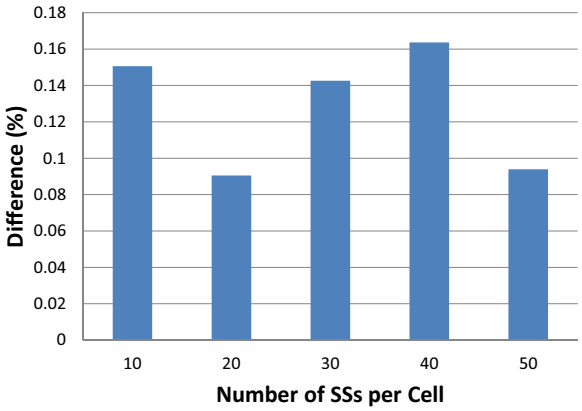
$$\frac{ILP\_Payoff - Greedy\_Payoff}{ILP\_Payoff} \times 100\%$$

This investigation gives us the numerical results representing the difference between the heuristic and optimal solutions. Fig. 5.3(b) and 5.4(b) are the investigation results for both environments. We can observe the difference of simulation results between ILP and greedy is at less than 0.2 % of ILP results. Fig. Thus, this confirms that our greedy algorithm can achieve nearly optimal solutions.

Fig 5.5 and 5.6 present the average delay and throughput for each SS in 2-cell and 3-cell environments, respectively. From the figures, we can observe that both ILP and greedy algorithm have similar results in these environment. Further, the average throughput in 2-cell environment is slight higher than then one in 3-cell environment due to less inter-cell interference. This reason also reflects lower average delay in 2-cell environment. We also investigate 7-cell environment. However, due to high computation complexity, ILP becomes intractable in 7-cell environment. We perform average delay and throughput for the greedy algorithm in Fig. 5.7. Due to stronger inter-cell interference, Fig. 5.7 shows lower throughput and higher delay comparing to 5.5 and 5.6. Further, it is worth noting that similar throughput is achieved in three tested environments. It shows that the QoS requirement is ensured in the proposed schemes.

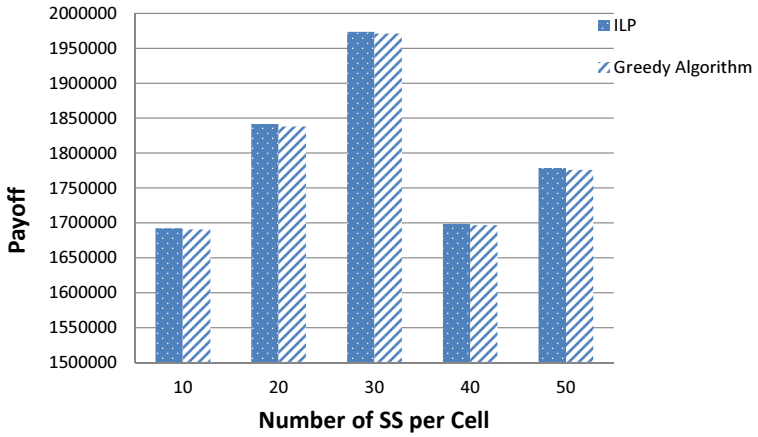


(a) Payoff Comparison

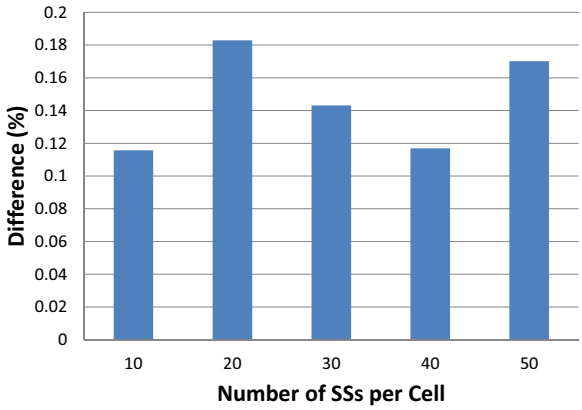


(b) Difference

Figure 5.3 Payoff for 2-cell Environment

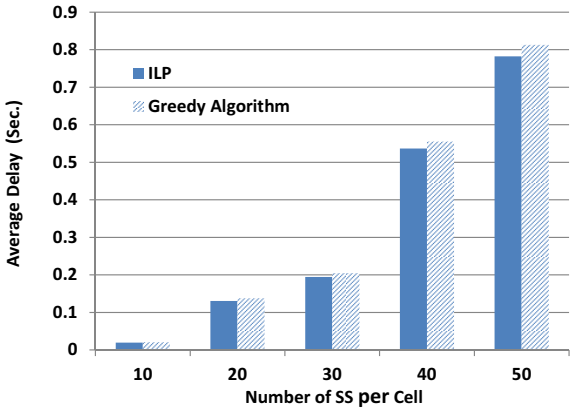


(a) Payoff Comparison

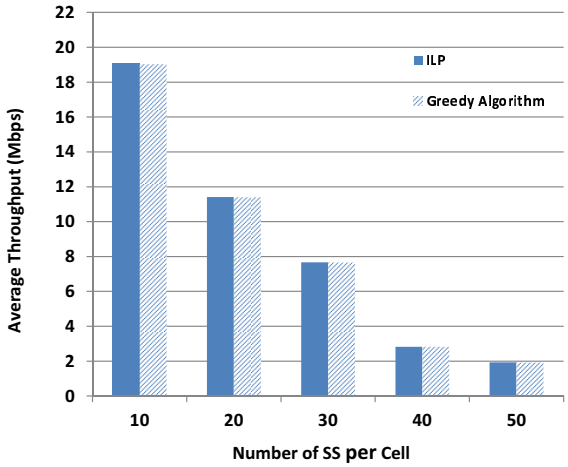


(b) Difference

Figure 5.4 Payoff for 3-cell Environment

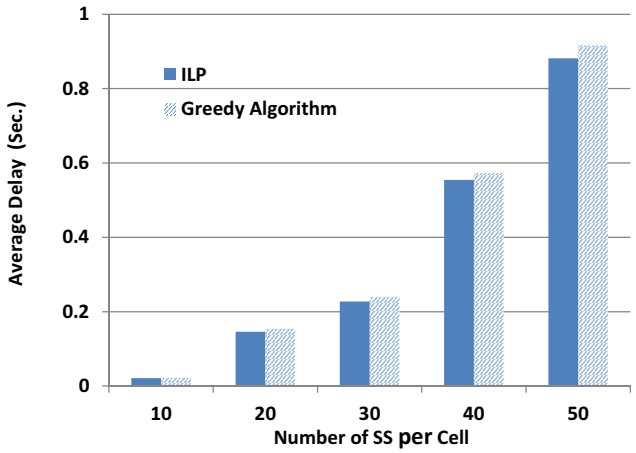


(a) Average Delay Comparison

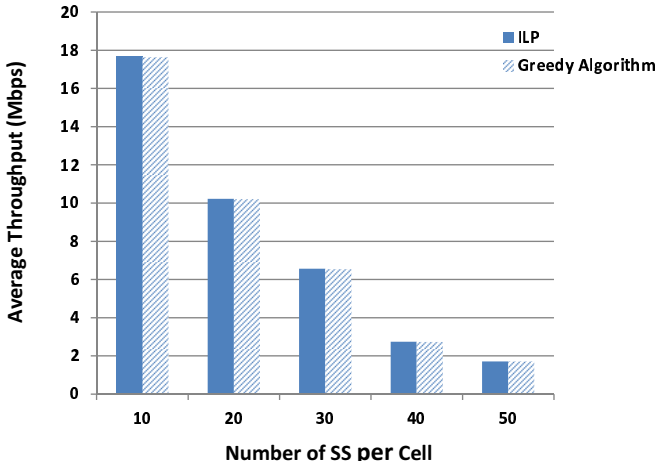


(b) Average Throughput Comparison

Figure 5.5 Average delay and throughput comparison for 2-cell Environment

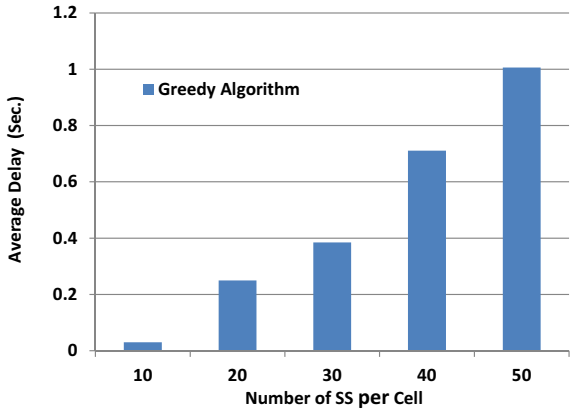


(a) Average Delay Comparison

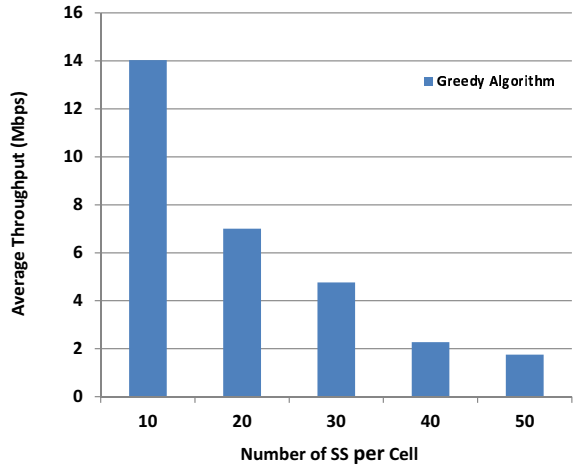


(b) Average Throughput Comparison

Figure 5.6 Average delay and throughput comparison for 3-cell Environment



(a) Average Delay Comparison



(b) Average Throughput Comparison

Figure 5.7 Average delay and throughput comparison for 7-cell Environment

We also implement two conventional objectives with 50 SSs and 2 cells in our simulation: maximizing system throughput (MAX-Throughput) and minimizing power consumption (MIN-Power). All constraints shown in Section 5 should also hold for these two objectives. We present the detail objective of these two objectives as below:

MIN-Power:

$$\sum_{i\in N}\sum_{s_j^i\in S_i}\sum_{b\in B}\sum_{p_l\in P} Y(s_j^i,b,p_l)\cdot\left(-P_c(s_j^i,b,p_l)\right)$$

MAX-Throughput:

$$\sum_{i\in N}\sum_{s_j^i\in S_i}\sum_{b\in B}\sum_{p_l\in P} Y(s_j^i,b,p_l)\cdot\left(R(s_j^i,b)\right)$$

We compare these two objectives to the proposed objective in terms of efficiency ratio. The efficiency ratio is defined as the ratio of the system throughput to the cost of power consumption. The simulation results of the three schemes are shown in Fig. 5.8. In the figure, we can observe that the proposed scheme results in higher efficiency ratio than MAX-Throughput. It is because in this objective, the BS only focuses on system throughput. It may leads to spend a high cost of power consumption for limited throughput improvement. On the other hand, MIN-Power leads to the lowest efficiency ratio because the BS tries to minimize the power consumption to just satisfy the QoS requirement of each SS. However, in our scheme, the BS allocates more bandwidth to the SS with good channel quality to boost up the system throughput with relatively small cost of power consumption.

5.8 Conclusion

In the paper, we focus on power allocation problem in dynamical fraction frequency reuse (DFFR). DFFR allows all available frequency sections to be utilized in each cell with dynamically changed transmission power corresponding to the current traffic demand in each cell. Due to this feature of DFFR, how to allocate transmission power in each cell is important and directly affects the system throughput.

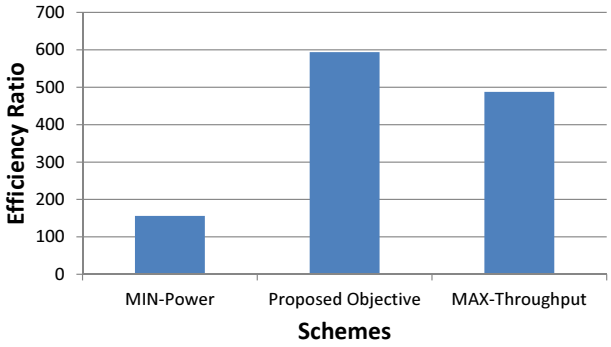


Figure 5.8 Scheme Comparison

We emphasis on the objective of allocating transmission power in each available frequency section such that the data transmission can be performed in the most energy-efficient way. Instead of minimizing the transmission power consumption, we target on maximizing the system throughput while minimizing the power consumption. We first formulate our problem by integer linear programming (ILP). Due to high computational complexity of ILP, we further propose a heuristic algorithm based on greedy approach. We implement our ILP formulation and greedy algorithm by CPLEX and JAVA, respectively. Our simulation results show that the greedy algorithm can achieve nearly optimal solutions.

Parameters	Description
i	i-th cell in the system
S^i	The set of SSs in cell i
S^i_r	The set of SSs serving real time traffic in cell i
S^i_{nr}	The set of SSs serving non-real time traffic in cell i
S^i_{be}	The set of SSs serving best effort traffic in cell i
B	The set of resource blocks in each cell
P	The set of transmission power level for each BS
s^i_j	j-th SS in cell i
b	b-th resource block
p_l	transmission power
$R(s^i_j, b)$	Throughput of s^i_j at RB b
$P_c(s^i_j, b, p_l)$	Power cost for s^i_j at RB b with transmission power p_l
$Y(s^i_j, b, p_l)$	Binary decision variable: 1: RB b is allocated to s^i_j with transmission power p_l. 0: Otherwise
P^i_{bs}	Power capacity for the BS in cell i
$R^{s^i_j}_{req}$	QoS requirement for s^i_j
$SINR^{s^i_j}_{req}$	SINR requirement for s^i_j
W	Maximum backoff window
T	Threshold for performing power allocation
Q_r	The current amount of data stored in queue
Q^f_r	The expected amount of data arrived until the next power allocation
$T^{max}_{i,j}$	The maximum delay requirement for s^i_j

Table 5.1 Parameters for ILP Formulation

Names	Values
Number of Cells	2, 3, 7
Number of SSs per cell	10, 20, 30, 40, 50
Number of BS per cell	1
BS service coverage	2km
SS distribution	Random
Random backoff interval	0 to 100 frames
Average deferred time	1 second
Modulation	QPSK, 16QAM, 64QAM
Frequency	2 GHz
Frame duration	10 ms
maximum backoff window	100

Table 5.2 Simulation Environment

Application	A1	A2	A3
Scheduling Class	C1	C2	C3
Minimum Traffic rate (bps)	2.05M	512 k	0
Maximum Sustained Rate (bps)	3.3M	25M	30K
Maximum delay (Sec.)	0.2	1*	1*
A1: Video Streaming A2: FTP A3: Web Browsing C1: Real Time C2: non-Real Time C3: Best Effort *The maximum delay requirement for FTP and web browsing only when the one for video streaming is ensured.			

Table 5.3 Traffic Parameters

MCS	**Modulation**	**Code Rate**	**SINR threshold** *[dB]*	**Efficiency** *[bits/ symbol]*
MCS 1	QPSK	1/12	-6.50	0.15
MCS 2	QPSK	1/9	-4.00	0.23
MCS 3	QPSK	1/6	-2.60	0.38
MCS 4	QPSK	1/3	-1.00	0.60
MCS 5	QPSK	1/2	1.00	0.88
MCS 6	QPSK	3/5	3.00	1.18
MCS 7	16QAM	1/3	6.60	1.48
MCS 8	16QAM	1/2	10.00	1.91
MCS 9	16QAM	3/5	11.40	2.41
MCS 10	64QAM	1/2	11.80	2.72
MCS 11	64QAM	1/2	13.00	3.32
MCS 12	64QAM	3/5	13.80	3.90
MCS 13	64QAM	3/4	15.60	4.52
MCS 14	64QAM	5/6	16.80	5.12
MCS 15	64QAM	11/12	17.60	5.55

Table 5.4 MCS (Modulation and Coding Schemes)

CHAPTER 6. Conclusion

In the next generation networks, providing quality of service (QoS) service is a fundamental feature. To achieve this feature, reservation based bandwidth allocation is adopted to ensure that the base station (BS) can guarantee the minimum QoS requirements for each subscriber station (SS). However, due to the characteristic of variable bit rate (VBR) traffic, it is very difficult to make an appropriate bandwidth reservation all the time. The bandwidth utilization may be degraded when the bandwidth is over-reserved. On the other hand, the QoS requirement may not be satisfied if the reserved bandwidth is less than the actual need. This thesis contains performance analysis in bandwidth request mechanism and the issue of bandwidth allocation in IEEE 802.16 networks. We propose both passive and active solutions to improve bandwidth utilization. At the end, we further investigate power consumption in wireless network and propose a joint optimization to achieve economical data transmission.

In this thesis, we first analyze two bandwidth request mechanisms in IEEE 802.16 networks. We provide mathematical models for each mechanism: unicast polling and contention resolution and perform performance analysis in terms of throughput and delay. We further propose two performance objectives: 1)minimizing delay with a fixed target throughput. 2) maximizing throughput while achieving a target delay requirement. We design two algorithms to help BS make scheduling decision to achieve each performance objective. The simulation results show that our algorithms can always help the BS make a better choice.

Due to the nature of bandwidth reservation, the bandwidth may not be utilized all the

time. In bandwidth recycling, we first investigate the percentage of unused bandwidth in a general network. We further propose a protocol named bandwidth recycling which allows the BS to schedule backup SSs to pick up the unused bandwidth. Based on the performance analysis of bandwidth recycling, we summarize the factors which affecting the performance and propose three additional algorithms to improve the performance. According to our simulation results, bandwidth recycling can averagely improve the system performance by 40%.

In additional to bandwidth recycling which utilizes the unused bandwidth, we further investigate the problem of minimizing unused bandwidth. We propose a game theoretic scheme to help the SS make bandwidth reservation with consideration of both QoS requirements and total bandwidth demand in the network. This scheme not only ensures QoS requirements in a heavily loaded network but also gives the flexibility of requesting more bandwidth when the network is lightly loaded. Our simulation and numerical results show a limited gap between the proposed scheme and optimal solutions derived from integer linear program.

With the consideration of power consumption, pursuing maximum system throughput might not be the best objective for networks. We investigate the issue of economical data transmissions considering both system throughput and power consumption. We propose a joint optimization with these two factors. With comparing to the existing schemes, based on the simulation results, the proposed scheme can reach the most economical data transmission.

In the future, we plan to continue to focus on the issue of economical data transmission with more practical condition such as jointly optimization of power, throughput and modulations. Moreover, heterogenous network environment becomes more popular in our daily life. it becomes more common that people can access multiple types of networks (e.g., WiFi, cellular network and WiMAX) at the same time. We are also interested in the issue of resource allocation in heterogenous networks as part of our future work.

Bibliography

[1] IEEE 802.16 WG,"IEEE Standard for Local and Metropolitan Area Network Part 16: Air Interface for Fixed Boardband Wireless Access Systems" IEEE Std 802.16-2004 p.1 - p.857

[2] Alexander Sayenko, Olli Alanen and Timo Hämäläinen, "On Contention Resolution Parameters for the IEEE 802.16 Base Station", GLOBECOM 2007, p4957 - 4962.

[3] Chun Nie, Muthaiah Venkatachalam and Xiangying Yang "Adaptive Polling Service for Next-Generation IEEE 802.16 WiMAX Networks", GLOBAECOM 2007, p4754-4758.

[4] Ben-Jye Chang, Chien-Ming Chou and Ying-Hsin Liang "Markov chain analysis of uplink subframe in polling-based WiMAX networks", Computer Communications 31 (2008), p.2381-2390

[20] Thomas G. Rpbertazzi "Computer Networks and Systems : Theory and Performance Evaluation." Springer-Verlag 1990.

[6] Yaser Pourmohammadi, Farshid Agharebparast, Mahmood R. Minhas, Hussein M. Alnuweiri and Victor C.M. Leung "Analytical Modeling of Contention-Based Bandwidth Requst Mechanism in IEEE 802.16 Wireless Networks" IEEE Transaction on Vehicular Technology, Vol. 57, No. 5 September 2008.

[7] Giuseppe Bianchi "Performance Analysis of the IEEE 802.11 Distributed Coordination Function", IEEE Journal on Selected Areas in Communications, Vol. 18, No. 3, March 2000.

[8] Jesús Delicado, Qiang Ni, Francisco M. Delicado and Luis Orozco-Barbosa "New Contention Resolution Schemes for WiMAX", WCNC 2009, p.1-6

[9] Hossam Fattah and Hussein Alnuweiri "Performance Evaluation of Contention-Based Access in IEEE 802.16 Networks with Subchannelization", ICC 2009, p.1-6

[10] Qiang Ni, Alexey Vinel, Yang Xiao, Andrey Turlikov and Tao Jiang "Investigation of Bandwidth Request Mechanisms under Point-to-MultiPoint Mode of WiMAX Networks" IEEE Communication Magazine, p132-138, May 2007.

[11] Robert M. Metcalfe and David R. Boggs, Ethernet: Distributed Packet Switching for Local Computer Networks, Communications of the ACM, vol. 19, no. 7, pp. 395 V 404, July 1976.

[12] M. Molina, P. Castelli and G. Foodies, "Web traffic modeling exploiting TCP connection temporal clustering through HTML-REDUCE", IEEE Network Magazine, vol. 14, no. 3, p.46-55, May 2000.

[13] Z. Sun, *et al.*, "Internet QoS and traffic modling", IEEE Proceedings Software, Special Issue on Performance Engineering, vol. 151, no. 5, p.248-255, October, 2004

[53] IEEE 802.16WG, "IEEE standard for local and metropolitan area networks part 16: Air interface for fixed and mobile broadband wireless access systems, Amendment 2," IEEE 802.16 Standard, December 2005.

[15] Jianhua He, Kun Yang and Ken Guild "A Dynamic Bandwidth Reservation Scheme for Hybrid IEEE 802.16 Wireless Networks" *ICC'08 p.2571-2575.*

[16] Kamal Gakhar, Mounir Achir and Annie Gravey, "Dynamic resource reservation in IEEE 802.16 broadband wireless networks", *IWQoS, 2006. p.140-148*

[17] J. Tao, F. Liu, Z. Zeng, and Z. Lin, Throughput enhancement in WiMax mesh networks using concurrent transmission, *In Proc. IEEE Int. Conf. Wireless Commun., Netw. Mobile Comput., 2005, p. 871V874.*

[18] Xiaofeng Bai, Abdallah Shami and Yinghua Ye "Robust QoS Control for Single Carrier PMP Mode IEEE 802.16 Systems",*IEEE TRANSACTIONS ON MOBILE COMPUTING, VOL. 7, NO. 4, APRIL 2008, p.416-429*

[19] Eun-Chan Park, Hwangnam Kim, Jae-Young Kim, Han-Seok Kim "Dynamic Bandwidth Request-Allocation Algorithm for Real-time Services in IEEE 802.16 Broadband Wireless Access Networks", *INFOCOM 2008,p.852 - 860*

[20] Thomas G. Robertazzi "Computer Networks and Systems:Theory and Performance Evaluation." *Springer-Verlag 1990*

[21] Kamal Gakhar, Mounir Achir and Annie Gravey, "How Many Traffic Classes Do We Need In WiMAX?," *WCNC 2007, p.3703-3708*

[22] Giuseppe Iazeolla, Pieter Kritzinger and Paolo Pileggi, "Modelling quality of service in IEEE 802.16 networks,"*SoftCOM 2008. p.130-134*

[23] Qualnet,http://www.scalable-networks.com/products/developer/new_in_45.php

[24] Frank H.P. Fitzek, Martin Reisslein, "MPEG–4 and H.263 Video Traces for Network Performance Evaluation", *IEEE Network, Vol. 15, No. 6, p.40-54, November/December 2001*

[25] Patrick Seeling, Martin Reisslein, and Beshan Kulapala, "Network Performance Evaluation Using Frame Size and Quality Traces of Single-Layer and Two-Layer Video: A Tutorial", *IEEE Communications Surveys and Tutorials, Vol. 6, No. 2 p.58-78, Third Quarter 2004*

[26] Geert Van der Auwera, Prasanth T. David, and Martin 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 Transactions on Broadcasting Vol. 54, No. 3 p.698-718 September 2008*

[54] K. Wongthavarawat and A. Ganz, "Packet scheudling for QoS support in IEEE 802.16 broadband wireless access systems" *International Journal of Communication Systems, vol. 16, pp 81-96, 2003*

[55] C. Cicconetti, L. Lenzini, E. Mingozzi and C. Eklund, "Quality of service suppoer in IEEE 802.16 networks" *IEEE Network, vol. 20, pp. 50-55, March/April 2006*

[56] G. Song, Y. Li, J. L. J. Cimini and H. Zheng, "Joint channel-aware and queue-aware data scheduling in multiple shared wireless channels" *In Proceedings of IEEE Wireless Communications and Networking Conference (WCNC), vol3, pp. 1939-1944, 2004*

[57] A. Sayenko, O. Alanen, J. Karhula, and T. Ḧaḿaḽainen, "Ensuring the qos requirments in 802.16 scheduling", *In Proceedings of the 9th ACM international symposium on Modeling analysis and simulation of wireless and mobile systems (MSWiM), pp. 108-117, 2006.*

[58] Q. Liu, X. Wang, and G. B. Giannakis, "A cross-layer scheduling algorithm with QoS support in wireless networks", *IEEE Transactions on Vehicular Technology, Vol. 55, pp. 839847, May 2006.*

[59] D. Niyato and E. Hossain, "Queue-aware uplink bandwidth allocation and rate control for polling service in IEEE 802.16 broadband wireless access networks", *IEEE Transactions on Mobile Computing, Vol. 5, No. 6, pp. 668679, 2006.*

[60] Eun-Chan Park, Hwanfnam Kim, Jae-Young Kim and Han Seok Kim, "Dynamic Bandwidth Request-Allocation Algorithm for Real-Time Services in IEEE 802.16 Broadband Wireless Access Networks", *INFOCOM '08, pp.852-860*

[62] J. Nash, "Equilibrium points in N-person games", *In Proceedings of National Academy of Science, Vol. 36, pp. 48-49, 1950*

[63] Mark Felegyhazi and Jean-Pierre Hubaux, "Game Theory in Wireless Networks: A Tutorial", *EPFL Technical report: LCA-REPORT-2006-002.*

[64] Dusit Niyato and Ekram Hossain, "QoS-aware bandwidth allocation and admission control in IEEE 802.16 broadband wireless access networks: a non-cooperative game theoretic approach", *Computer Communication Vol. 51, pp. 3305-3321, 2007.*

[65] Hui Zhang and Xuming Fang, "A Pricing and Game Theory-based Call Admission Control Scheme for CDMA Systems", *Wireless Communication, Networking and Mobile Computing, pp. 6493-6496, 2007.*

[66] Angelos N. Rouskas, Anastasios A. Kikilis and and Stilianos S. Ratsiatos, "A game theoretical formulation of integrated admission control and pricing in wireless networks", *European Journal of Operational Research, Vol. 191, pp. 11751188, 2008*

[67] Shamik Sengupta, Mainak Chatterjee and Kevin Kwiat, "A Game Theoretic Framework for Power Control in Wireless Sensor Networks", *IEEE Transactions on Computers, 2009, appeared online: http://doi.ieeecomputersociety.org/10.1109/TC.2009.82*

[68] D. Niyato, E. Hossain, M.M. Rashid and V.K. Bhargava, "Wireless sensor networks with energy harvesting technologies: a game-theoretic approach to optimal energy management" *IEEE Wireless Communication, VOl. 14, Issue 4 pp. 90-96, 2007*

[69] Beibei Wang, Zhu Han, and K.J.R. Liu, "Distributed Relay Selection and Power Control for Multiuser Cooperative Communication Networks Using Stackelberg Game" *IEEE Transactions on Mobile Computing, Vol. 8, Issue 7 pp 975-990, 2009*

[70] D. Niyato, E. Hossain and Zhu Han, "Dynamics of Multiple-Seller and Multiple-Buyer Spectrum Trading in Cognitive Radio Networks: A Game-Theoretic Modeling Approach" *IEEE Transactions on Mobile Computing, Vol. 8, Issue 8, pp. 1009-1022, 2009*

[71] Lin Gao, Xinbing Wang and Youyun Xu, "Multiradio Channel Allocation in Multihop Wireless Networks", *IEEE Transactions on Mobile Computing, Vol. 8, Issue 11, pp. 1454-1468, 2009*

[74] J. Neumann and O. Morgenstern, "Theory of Games and Economoc Behavior." *Princeton University Press, 1944.*

[75] Mingbo Xiao, Ness B. Shroff and Edqin K. P. Chong, " Utility-Based Power Control in Celleular Wireless Systems" *INFOCOM'01*, p.412-421.

[78] Chakchi So-In, Raj Jain and Abdel-Karim Tamimi,"Scheudling in IEEE 802.16e Mobile WiMAX Networks: Key Issues and a Survey", *J'SAC '09*,p.156-171

[80] S.A. Filin, S.N. Moiseev and M.S. Kondakov, "Fast and Efficient QoS-Guaranteed Adaptive Transmission Algorithm in the Mobile WiMAX System", *IEEE Transactions on Vechicular Technology, Vol. 57, Issue 6, pp. 3477-3487 2008*

[81] Y. Ahmet Sekercioglua, Milosh Ivanovicha and Alper Yeginb, "A survey of MAC based QoS implementations for WiMAX networks", *Computer Networks Vol. 53, Issue 14, pp. 2517-2536 2009*

[49] Alexander L. Stolyar and Harish Viswanathan, "Self-organizing Dynamic Franctional Frequency Reuse in OFDMA Systems",*INFOCOM'08*,p.691 - 699.

[50] Alexander L. Stolyar and Harish Viswanathan, "Self-organizing Dynamic Franctional Frequency Reuse for Best-Effort Traffic Through Distributed Inter-cell Coodination",*INFOCOM'09*,p.1287 - 1295.

[51] Eduard Jorswieck and Rami Mochaourab, "Power Control Game in Protected and SHared Bands manipuability of Nash Equilibrium",*GameNet'09*,p.428 - 437.

[52] Z. Abichar, J. M. Chang, and C. Y. Hsu, " WiMAX or LTE: Who will Lead the Broadband Mobile Internet?", IEEE IT Professional, Volume 12, number 3, May, 2010, pp. 26-32

[53] IEEE 802.16WG, "IEEE standard for local and metropolitan area networks part 16: Air interface for fixed and mobile broadband wireless access systems, Amendment 2," *IEEE 802.16 Standard*, December 2005.

[54] K. Wongthavarawat and A. Ganz, "Packet scheudling for QoS support in IEEE 802.16 broadband wireless access systems" *International Journal of Communication Systems, vol. 16, pp 81-96, 2003*

[55] C. Cicconetti, L. Lenzini, E. Mingozzi and C. Eklund, "Quality of service suppoer in IEEE 802.16 networks" *IEEE Network, vol. 20, pp. 50-55, March/April 2006*

[56] G. Song, Y. Li, J. L. J. Cimini and H. Zheng, "Joint channel-aware and queue-aware data scheduling in multiple shared wireless channels" *In Proceedings of IEEE Wireless Communications and Networking Conference (WCNC), vol3, pp. 1939-1944, 2004*

[57] A. Sayenko, O. Alanen, J. Karhula, and T. Ḧaḿaḽainen, "Ensuring the qos requirments in 802.16 scheduling", *In Proceedings of the 9th ACM international symposium on Modeling analysis and simulation of wireless and mobile systems (MSWiM), pp. 108-117, 2006.*

[58] Q. Liu, X. Wang, and G. B. Giannakis, "A cross-layer scheduling algorithm with QoS support in wireless networks", *IEEE Transactions on Vehicular Technology, Vol. 55, pp. 839847, May 2006.*

[59] D. Niyato and E. Hossain, "Queue-aware uplink bandwidth allocation and rate control for polling service in IEEE 802.16 broadband wireless access networks", *IEEE Transactions on Mobile Computing, Vol. 5, No. 6, pp. 668679, 2006.*

[60] Eun-Chan Park, Hwanfnam Kim, Jae-Young Kim and Han Seok Kim, "Dynamic Bandwidth Request-Allocation Algorithm for Real-Time Services in IEEE 802.16 Broadband Wireless Access Networks", *INFOCOM '08, pp.852-860*

[61] David Chuck and J. Morris Chang, "Bandwidth Recycling in IEEE 802.16 Networks", *IEEE Transactions on Mobile Computing, Vol. 9, No. 10 pp. 1451-1464, 2010*

[62] J. Nash, "Equilibrium points in N-person games", *In Proceedings of National Academy of Science, Vol. 36, pp. 48-49, 1950*

[63] Mark Felegyhazi and Jean-Pierre Hubaux, "Game Theory in Wireless Networks: A Tutorial", *EPFL Technical report: LCA-REPORT-2006-002.*

[64] Dusit Niyato and Ekram Hossain, "QoS-aware bandwidth allocation and admission control in IEEE 802.16 broadband wireless access networks: a non-cooperative game theoretic approach", *Computer Communication Vol. 51, pp. 3305-3321, 2007.*

[65] Hui Zhang and Xuming Fang, "A Pricing and Game Theory-based Call Admission Control Scheme for CDMA Systems", *Wireless Communication, Networking and Mobile Computing, pp. 6493-6496, 2007.*

[66] Angelos N. Rouskas, Anastasios A. Kikilis and and Stilianos S. Ratsiatos, "A game theoretical formulation of integrated admission control and pricing in wireless networks", *European Journal of Operational Research, Vol. 191, pp. 11751188, 2008*

[67] Shamik Sengupta, Mainak Chatterjee and Kevin Kwiat, "A Game Theoretic Framework for Power Control in Wireless Sensor Networks", *IEEE Transactions on Computers, 2009, appeared online: http://doi.ieeecomputersociety.org/10.1109/TC.2009.82*

[68] D. Niyato, E. Hossain, M.M. Rashid and V.K. Bhargava, "Wireless sensor networks with energy harvesting technologies: a game-theoretic approach to optimal energy management" *IEEE Wireless Communication, VOl. 14, Issue 4 pp. 90-96, 2007*

[69] Beibei Wang, Zhu Han, and K.J.R. Liu, "Distributed Relay Selection and Power Control for Multiuser Cooperative Communication Networks Using Stackelberg Game" *IEEE Transactions on Mobile Computing, Vol. 8, Issue 7 pp 975-990, 2009*

[70] D. Niyato, E. Hossain and Zhu Han, "Dynamics of Multiple-Seller and Multiple-Buyer Spectrum Trading in Cognitive Radio Networks: A Game-Theoretic Modeling Approach" *IEEE Transactions on Mobile Computing, Vol. 8, Issue 8, pp. 1009-1022, 2009*

[71] Lin Gao, Xinbing Wang and Youyun Xu, "Multiradio Channel Allocation in Multihop Wireless Networks", *IEEE Transactions on Mobile Computing, Vol. 8, Issue 11, pp. 1454-1468, 2009*

[72] Zhu, Kun; Niyato, Dusit; Wang, Ping; , "Network Selection in Heterogeneous Wireless Networks: Evolution with Incomplete Information," Wireless Communications and Networking Conference (WCNC), 2010 IEEE , vol., no., pp.1-6, 18-21 April 2010 doi: 10.1109/WCNC.2010.5506371

[73] Xingwei Liua, Xuming Fangb, Xu Chena and Xuesong Penga, "A bidding model and cooperative game-based vertical handoff decision algorithm" Journal of Network and Computer Applications, doi:10.1016/j.jnca.2011.01.012

[74] J. Neumann and O. Morgenstern, "Theory of Games and Economoc Behavior." *Princeton University Press, 1944.*

[75] Mingbo Xiao, Ness B. Shroff and Edqin K. P. Chong, " Utility-Based Power Control in Celleular Wireless Systems" *INFOCOM'01*, p.412-421.

[76] Scott Shenker, "Fundamental Design Issues for the Furture Internet", *IEEE Journal on Selected Area in Communications, Vol. 13, No. 7, pp. 1176-1188 1995*

[77] George D. Sramoulis, Dimitrios Kalopsilalis, Anna Kyrikoglou and Costas Courcoubetis, "Efficient Agent-based Negotiation for Telecommunication Services", *Globecom'99, pp 1989 - 1996*

[78] Chakchi So-In, Raj Jain and Abdel-Karim Tamimi,"Scheudling in IEEE 802.16e Mobile WiMAX Networks: Key Issues and a Survey", *J'SAC '09*,p.156-171

[79] Gaoning He, Merouane Debbah, and Eitan Altman, "A Bayesian Game-Theoretic Approach for Distributed Resource Allocation in Fading Multiple Access Channels", *EURASIP Journal on Wireless Communications and Networking Volume 2010 (2010)*

[80] S.A. Filin, S.N. Moiseev and M.S. Kondakov, "Fast and Efficient QoS-Guaranteed Adaptive Transmission Algorithm in the Mobile WiMAX System", *IEEE Transactions on Vechicular Technology, Vol. 57, Issue 6, pp. 3477-3487 2008*

[81] Y. Ahmet Sekercioglua, Milosh Ivanovicha and Alper Yeginb, "A survey of MAC based QoS implementations for WiMAX networks", *Computer Networks Vol. 53, Issue 14, pp. 2517-2536 2009*

[82] AMPL, http://www.ampl.com/

[83] Piro, G.; Grieco, L.A.; Boggia, G.; Capozzi, F.; Camarda, P.; , "Simulating LTE Cellular Systems: An Open-Source Framework," Vehicular Technology, IEEE Transactions on , vol.60, no.2, pp.498-513, Feb. 2011 doi: 10.1109/TVT.2010.2091660

[84] Xu Fangmin; , "Fractional Frequency Reuse (FFR) and FFR-Based Scheduling in OFDMA Systems," Multimedia Technology (ICMT), 2010 International Conference on , vol., no., pp.1-4, 29-31 Oct. 2010 doi: 10.1109/ICMULT.2010.5629810

[85] Chang, R.Y.; Zhifeng Tao; Jinyun Zhang; Kuo, C.-C.J.; , "A Graph Approach to Dynamic Fractional Frequency Reuse (FFR) in Multi-Cell OFDMA Networks," Communications, 2009. ICC '09. IEEE International Conference on , vol., no., pp.1-6, 14-18 June 2009 doi: 10.1109/ICC.2009.5198612

[86] Chi-Yao Hong,Ai-Chun Pang and Pi-Cheng Hsiu"Approximation Algorithms for a Link Scheduling Problem in Wireless Relay Networks with QoS Guarantee", IEEE TRANSACTIONS ON MOBILE COMPUTING, VOL. 9, NO. 12, DECEMBER 2010.

[87] Hung-Cheng Shih and Kuochen Wang"A QoS-Guaranteed Energy-Efficient Packet Scheduling Algorithm for WiMax Mobile Devices" WIRELESS ALGORITHMS, SYSTEMS, AND APPLICATIONS Lecture Notes in Computer Science, 2010, Volume 6221/2010, 75-79

[88] Stolyar, A.L.; Viswanathan, H.; , "Self-Organizing Dynamic Fractional Frequency Reuse in OFDMA Systems," INFOCOM 2008. The 27th Conference on Computer Communications. IEEE , vol., no., pp.691-699, 13-18 April 2008 doi: 10.1109/INFOCOM.2008.119

[89] Stolyar, A.L.; Viswanathan, H.; , "Self-Organizing Dynamic Fractional Frequency Reuse for Best-Effort Traffic through Distributed Inter-Cell Coordination," INFOCOM 2009, IEEE , vol., no., pp.1287-1295, 19-25 April 2009 doi: 10.1109/INFCOM.2009.5062043

[90] Tara Ali-Yahiya and Haloma Chaouchi, "Fractional Frequency Reuse for Hierarchical Resourse Allocation in Mobile Networks", EURASIP Journal on Wireless Communication and Networking, Vol. 2010 Article ID 363065.

[91] Ali, S.H.; Leung, V.C.M.; , "Dynamic frequency allocation in fractional frequency reused OFDMA networks," Wireless Communications, IEEE Transactions on , vol.8, no.8, pp.4286-4295, August 2009 doi: 10.1109/TWC.2009.081146

[92] Peng Wang; Chunhui Liu; Mathar, R.; , "Dynamic fractional frequency reused proportional fair in time and frequency scheduling in OFDMA networks," Wireless Communication Systems (ISWCS), 2011 8th International Symposium on , vol., no., pp.745-749, 6-9 Nov. 2011

[93] Assaad, M.; , "Optimal Fractional Frequency Reuse (FFR) in Multicellular OFDMA System," Vehicular Technology Conference, 2008. VTC 2008-Fall. IEEE 68th , vol., no., pp.1-5, 21-24 Sept. 2008 doi: 10.1109/VETECF.2008.381

[94] Donghee Kim, Jae Young Ahn, and Hojoon Kim, "Downlink Transmit Power Allocation in Soft Fractional Frequency Reuse Systems" ETRI Journal, Volume 33, Number 1, February 2011

[95] Hamouda, S.; Choongil Yeh; Jihyung Kim; Shin Wooram; Dong Seung Kwon; , "Dynamic hard Fractional Frequency Reuse for mobile WiMAX," Pervasive Computing and Communications, 2009. PerCom 2009. IEEE International Conference on , vol., no., pp.1-6, 9-13 March 2009 doi: 10.1109/PERCOM.2009.4912869

[96] Novlan, T.D.; Ganti, R.K.; Ghosh, A.; Andrews, J.G.; , "Analytical Evaluation of Fractional Frequency Reuse for OFDMA Cellular Networks," Wireless Communications, IEEE Transactions on , vol.10, no.12, pp.4294-4305, December 2011

[97] David Chuck, K.Y. Chen and J. M. Chang, A Comprehensive Analysis of Bandwidth Request Mechanisms in IEEE 802.16 Networks, IEEE Transactions on Vehicular Technology, Vol. 59, No. 4,p.2046-2056, May 2010

[98] CPLEX, http://www-01.ibm.com/software/integration/optimization/cplex-optimizer/

[99] Forsk atoll - global RF planning solution

[100] Lopez-Perez, D.; Ladanyi, A.; Juttner, A.; Rivano, H.; Jie Zhang; , "Optimization method for the joint allocation of modulation schemes, coding rates, resource blocks and power in self-organizing LTE networks," INFOCOM, 2011 Proceedings IEEE , vol., no., pp.111-115, 10-15 April 2011 doi: 10.1109/INFCOM.2011.5934888