QoS Scheduling in IEEE 802.16 Broadband Wireless Access Networks

by

Fen Hou

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Doctor of Philosophy
in
Electrical and Computer Engineering

Waterloo, Ontario, Canada, 2008

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Abstract

With the exploding increase of mobile users and the release of new wireless applications, the high bandwidth requirement has been taking as a main concern for the design and development of the wireless techniques. There is no doubt that broadband wireless access with the support of heterogeneous kinds of applications is the trend in the next generation wireless networks. As a promising broadband wireless access standard, IEEE 802.16 has attracted extensive attentions from both industry and academia due to its high data rate and the inherent media access control (MAC) mechanism, which takes the service differentiation and quality of service (QoS) provisioning into account.

To achieve service differentiation and QoS satisfaction for heterogenous applications is a very complicated issue. It refers to many fields, such as connection admission control (CAC), congestion control, routing algorithm, MAC protocol, and scheduling scheme. Among these fields, packet scheduling plays one of the most important roles in fulfilling service differentiation and QoS provisioning. It decides the order of packet transmissions, and provides mechanisms for the resource allocation and multiplexing at the packet level to ensure that different types of applications meet their service requirements and the network maintains a high resource utilization.

In this thesis, we focus on the packet scheduling for difficult types of services in IEEE 802.16 networks, where unicast and mulitcast scheduling are investigated. For unicast scheduling, two types of services are considered: non-real-time polling service (nrtPS) and best effort (BE) service. We propose a flexible and efficient resource allocation and scheduling framework for nrtPS applications to achieve a tradeoff between the delivery delay and resource utilization, where automatic repeat request (ARQ) mechanisms and the adaptive modulation and coding (AMC) technique are jointly considered. For BE service, considering the heterogeneity of subscriber stations (SSs) in IEEE 802.16

networks, we propose the weighted proportional fairness scheduling scheme to achieve the flexible scheduling and resource allocation among SSs based on their traffic demands/patterns. For multicast scheduling, a cooperative multicast scheduling is proposed to achieve high throughput and reliable transmission. By using the two-phase transmission model to exploit the spatial diversity gain in the multicast scenario, the proposed scheduling scheme can significantly improve the throughput not only for all multicast groups, but also for each group member. Analytical models are developed to investigate the performance of the proposed schemes in terms of some important performance measurements, such as throughput, resource utilization, and service probability. Extensive simulations are conducted to illustrate the efficient of the proposed schemes and the accuracy of the analytical models. The research work should provide meaningful guidelines for the system design and the selection of operational parameters, such as the number of TV channels supported by the network, the achieved video quality of each SS in the network, and the setting of weights for SSs under different BE traffic demands.

Acknowledgments

I would like to express my deepest and sincerest gratitude to my supervisors: Professor Pin-Han Ho and Professor Xuemin (Sherman) Shen for their support and guidance during my study at the University of Waterloo. They have provided me with a motivating, enthusiastic, and critical research atmosphere. It is my honor to pursue my Ph.D. study under their supervision. They guide me not only the way to do research but also the attitude in life, which are of great benefit to me forever.

My extreme appreciation goes to my thesis committee members: Professor Sagar Naik, Professor Guang Gong, Professor Jun Cai, and Professor Min Song. They contributed their precious time to read my thesis, and provided valuable suggestions and comments that helped to improve the quality of this thesis.

My special thanks are due to Ms. Wei Song, Ms. Ping Wang, Ms. Lin Cai, Ms. Bin Lin, and Ms. Xiaojing Meng for their friendship during the last few years, which makes my life abroad easier and bring me many warm and happy memories.

I wish to thank Dr. Jon W. Mark, the director of the broadband communications research (BBCR) group. His solid knowledge and insightful comments in the group meetings have been of great value for me. I also wish to thank Dr. Xinhua Ling, Mr. Stanley Liu, Ms. Mehri Mehrjoo, Mr. James She, and other colleagues for their friendly help, stimulating discussions, and valuable suggestions. It is my great pleasure to be a member of BBCR group, a wonderful research team and a large warm family.

Many thanks go to the administrative staff: Ms. Wendy Boles, Ms. Lisa Hendel, and Ms. Karen Schooley for their kindness, patience, and help.

I am greatly indebted to my parents for their love and my husband for his supporting, understanding, and encouraging, which are indispensable for the completion of my Ph.D. study.

To my dear parents and husband

Contents

List of Tables

List of Figures

Chapter 1

Introduction

With the exploding increase of mobile users and the release of more and more wireless applications, the requirement for bandwidth becomes a more and more important issue [1–3]. From the first generation (1G) cellular networks to the second generation (2G) systems, up to the third generation (3G) and beyond 3G systems, bandwidth is a main concern when a system is upgraded. Following this trend, some leading standard associations, such as the Institute of Electrical and Electronics Engineering (IEEE) and the European Telecommunications Standard Institute (ETSI), set up working groups focusing on the standardization, development, and deployment of broadband communications.

1.1 Broadband Wireless Networks

Mobile and wireless communication techniques have experienced a tremendous development in the last few years. The cellular system is upgraded from 2G to 2.5G, 3G and beyond with the purpose of providing a higher data rate. High bandwidth and quality of service (QoS) provisioning are the critical goals in the development of communication techniques. The federal communications commission (FCC) defines broadband as a service or connection allowing considerable information to be transmitted with a capacity of at least 200 Kbps in at least one direction. Digital subscriber line (DSL) and cable modem (CM) are two popular approaches to provide the broadband access through a wired connection. DSL is a wiredline transmission technology that can provide a faster data transmission rate over traditional cooper telephone lines already installed at homes and business offices. DSL-based broadband access provides transmission rates ranging from several hundred Kbps to millions of bits per second. CM service enables cable operators to provide broadband data transmissions with the rate of 1.5 Mbps or more by using hybrid fiber coax (HFC) cable networks, which have been mainly used in the past to deliver TV signals to home.

Compared with wired communications, wireless communications has many unique features. First, it provides a broad geographical coverage with cost-efficient infrastructure deployment. Especially, for the locations wired line can not reach to. Secondly, it provides freedom not only for service providers but also for users. Wireless communication enables patrons to bring their own laptops, therefore, reducing the cost of owning many personal computers. Meanwhile, users can move their laptops. Broadband wireless networks are critical in developing and releasing new wireless applications. To extend the broadband access to wireless scenarios, IEEE 802.11 standards including IEEE 802.11b, and its successors IEEE 802.11a and IEEE 802.11g are designed for wireless local area network (WLAN) to provide wireless broadband access services. In early 2003, a boom in the sales of IEEE 802.11b WLAN products, also known as Wireless Fidelity (Wi-Fi), started to take place. IEEE 802.11b uses the unlicensed industrial 2.4 GHz frequency band to enable multiple computers and portable devices to connect to one or more access points, thus access to the Internet. It allows for the wireless transmission rate of approximately 11 Mbps of raw data at indoor distances from several meters to several hundred feet and outdoor distance of several to tens of miles. The IEEE 802.11a uses 5 GHz frequency band and can handle 54 Mbps at a typically shorter distance. In addition, wireless personal area network (WPAN) is designed for communications within a very short range (about 10 meters) with a high data rate. It could serve to interconnect all the computing and communicating devices around an individual person's workspace.

To further utilize the attractive features of the broad geography coverage with cost-efficient deployment, wireless broadband access is extended to metropolitan area network (MAN). The IEEE sets up the 802.16 working group and published its first IEEE 802.16 standard at the end of 2001. Later, several amendments and extensions are released to extend the frequency band and enhance the support for high data rate and mobility. The IEEE 802.16a standard aims to have a coverage of up to about 30 miles with data rate of up to 75 Mbps. Due to the capability of providing high data rate, service differentiation, and QoS satisfaction, the IEEE 802.16 standard attracts much more attentions from both academia and industry [4, 5]. Major service providers and equipment manufacturers, such as Intel, Nokia, and Fujitsu, have recently indicated that they support worldwide interoperability for microwave access (WiMAX), which promotes the IEEE 802.16 standard for broadband wireless access (BWA) networks. As a broadband wireless access technique, the IEEE 802.16 network is a promising and cost-effective alternative to CM and DSL services. Therefore, how to fulfill the QoS satisfaction and service differentiation in IEEE 802.16 networks is one of the most important and open issues.

1.2 QoS Provisioning

QoS refers to the capability of a network to provide different priorities to different applications, users, and data flows, or to guarantee a certain level of performance or behavior to a data flow. The specific parameters defining QoS depend on the types of applications and user requirements. Generally, QoS is mainly measured in terms of delay, jitter, bandwidth, and packet loss. How to fulfill QoS provisioning is always a key research issue in the field of communications. Dating from eighties in the 20th century, Internet appeared as one of the most important breakpoints in the field of communications. It provides many attractive features and packet-oriented transmissions. Since the Internet is designed originally to provide a same-for-all best-effort packet delivery service, it falls lack of any QoS guarantee in terms of packet delay, packet loss rate, and throughput. However, with the development of the Internet, QoS provisioning becomes a key challenge needed to be addressed. It is critical for supporting heterogenous kinds of services, such as real-time traffic or multimedia services [6, 7]. Thus, it is one of the key issues for both network service provider and end users. The Internet Engineering Task Force (IETF) has proposed some service models and mechanisms to fulfill QoS provisioning. Two well-known mechanisms to provide QoS support over the Internet are the integrated services (IntServ) and the differentiated services (DiffServ). The IntServ mechanism can provide a hard QoS guarantee by using the resource reservation protocol, while the Diffserv provides a class-based QoS satisfaction [8–10].

With the development of wireless technique and the popularity of wireless network environment, more and more mobile users connect to the Internet for multiple kinds of services, such as sending email, downloading files, browsing web sites, and making calls. In the future, broadband wireless networks will be an integral part of the global communication infrastructure. Meanwhile, with the rapid growth of the amount of wireless end users and the increasing demands for multimedia applications, there is no doubt that future wireless networks will provide services for heterogeneous classes of traffic with different QoS requirements. Compared with the wired networks, wireless networks exhibit some distinct features: (1)time-varying and location-dependent wireless channels; (2)high packet loss rate and burst error; (3)limited bandwidth; and (4)power constraint for mobile users. These characteristics pose new challenges for providing QoS in the wireless network environments. For instance, due to radio propagation characteristics, the achievable wireless channel capacity may be substantially degraded by impairments such as large-scale path loss and small-scale fading resulting from multipath time delay spread. In this case, how to mitigate the negative impacts of channel fading and improve the resource utilization is one of the most important issues.

QoS provisioning in wireless networks is a complicated issue involving several aspects and different layers. From the aspect of the network layer, QoS routing aims to find an optimal-cost path from a source to a destination subject to one or multiple QoS constraints [11–13]. From the aspect of the transportation layer, connection admission control (CAC) provides mechanisms for bandwidth allocation at the connection level. It determines whether a connection request can be accepted or not, thus ensuring some high priority connections have chance to achieve a high QoS service by rejecting some low-priority connections [14–16]. Congestion control schemes shape the traffic injecting into a network to ensure the network is not overloaded [17, 18]. From the aspect of the media access control (MAC) layer, packet scheduling provides mechanisms for the bandwidth allocation and multiplexing at the packet level. It plays a key role in providing service differentiation and QoS satisfaction [19–21]. In this thesis, we focus on the QoS scheduling for different types of services and communications in IEEE 802.16 networks.

1.3 Motivation and Objectives

As a promising broadband wireless access standard, IEEE 802.16 defines the objectives of providing service differentiation and QoS satisfaction. However, it does not specify any specific scheme in terms of scheduling, CAC, and congestion control to fulfill the objectives. Packet scheduling deals with the allocation of network resources among different users and different types of service flows. It is critical in QoS provisioning. Many previous research work focused on the conventional networks, such as 3G cellular systems where all end users are individual cell phones or handsets, can not be extended to IEEE 802.16 networks due to their unique features, such as heterogeneous traffic demands of subscriber stations (SSs) and the adaptive modulation and coding (AMC) technique. Proportional fairness scheduling in [22, 23] can achieve a good tradeoff between the throughput and fairness, which is efficient when the traffic demands of users are homogeneous. However, it is not suitable for IEEE 802.16 networks due to the potential heterogeneity among SSs in terms of average channel conditions and traffic demands. On the other hand, previous research work on IEEE 802.16 networks is mainly focused on real-time polling service (rtPS) and BE service. However, non-real-time polling service (nrtPS) flows account for considerable amount of traffic and deserves more investigation. In addition, many existing studies are simulation-based research. Developing analytical models are practically important, which can be used in the performance analysis of resource management and provide a meaningful guideline on the design of an efficient admission control scheme and the implementation of effective upper-layer network protocols. Therefore, how to design and analyze efficient scheduling schemes for nrtPS and BE services to achieve QoS satisfaction and remain a high radio resource utilization is a critical issue to fulfill the objectives specified in the IEEE 802.16 standard.

With the development in broadband wireless access technique and scalable video technologies, multimedia services, such as Internet protocol television (IPTV), video conference, and network gaming, are expected to be killer applications in the next generation wireless metropolitan area networks (WMANs) [24, 25]. These multimedia services are one-to-many applications. Due to the broadcast nature of wireless communications, multicast transmission is an efficient way to provide services to multiple users simultaneously. It is, therefore, adopted to support these emerging multimedia services. However, how to design an efficient multicast scheduling scheme for supporting multimedia services is a challenging issue. In a multicast scenario, different multicast groups have different sets of group members distributed at different locations. Generally, group members experience different channel conditions. Therefore, how to select multicast group for service and how to efficiently multicast data to all group members in the selected multicast group are two key issues. If the selection of multicast group is based on the best channel condition among all members in a multicast group, the achieved throughput may not be high if most of the other group members experience bad channel conditions. If the selection of multicast group is based on the sum of channel conditions of all group members, it may lead to unfairness because multicast groups close to the base station (BS) usually have good channel conditions, and thus are more likely to be scheduled and dominate the bandwidth consumption. Meanwhile, If the transmission rate is set too high, some group members with bad channel conditions may not be able to successfully decode the data. On the contrary, if the rate is determined based on the group member with the worst channel condition, the wireless resource would be underutilized since some group members with good channel conditions can support a higher data rate. Therefore, how to deal with the diverse channel conditions of group members in a multicast scenario and design an efficient multicast scheduling scheme is critical and deserves more investigation.

Thus, motivated, we address the packet scheduling in both unicast and multicast scenarios. For unicast scheduling, we focus on nrtPS and BE service for achieving a flexible and efficient scheduling while considering the unique features of IEEE 802.16 networks and some advance techniques specified in the IEEE 802.16 standard, such as AMC technique at the physical (PHY) layer and automatic repeat request (ARQ) mechanisms at the MAC layer. For the multicast scheduling, we propose an efficient multicast scheduling scheme for achieving the high throughput and resource utilization for all group members in both bad and good channel conditions.

1.4 Main Contributions

This thesis focuses on the packet scheduling in IEEE 802.16 networks. Its main contributions are listed as follows:

- An efficient yet simple resource allocation and scheduling framework is proposed for nrtPS flows to achieve a flexible tradeoff between the delivery delay and resource utilization while maintaining their minimum bandwidth requirements.

- An analytical model is developed for parameter manipulation in the proposed framework. By jointly considering ARQ mechanisms at the MAC layer in the analytical model, we analyze the achieved goodput of each SS, which is a more important performance measurement than assigned bandwidth from the receiver's point of view. Meanwhile, AMC technique at the PHY layer is taken into account as well, and some important performance metrics, such as service probability, resource utilization, and fairness, are investigated for evaluating the performance and efficiency of the proposed framework.

- For BE service in IEEE 802.16 networks, the weighted proportional fair (WPF) scheduling scheme is proposed for achieving a flexible resource allocation among SSs considering their different traffic demand.

- An analytical model is developed to investigate the performance of WPF in terms of service probability, spectral efficiency, throughput, and fairness. By using the analytical model, we evaluate the impact of the weight of each SS and channel conditions on these performance metrics, and quantify the relation between weights, channel conditions, and these performance metrics.

- A cooperative multicast scheduling (CMS) scheme is proposed to achieve efficient multicast transmissions for supporting multimedia services. By selecting the multicast group based on the normalized average channel condition, CMS can achieve a good fairness in terms of channel access. Furthermore, by exploiting the spatial diversity among multiple users and the cooperative transmissions, CMS can provide a significant enhancement of throughput not only for the SSs with good channel conditions, but also for the SSs with bad channel conditions.

- An analytical model is developed to evaluate the performance of CMS in terms of service probability of each multicast group (MGroup) and achieved throughput of each group member and the whole network. The analysis results can provide a meaningful guideline for the system design, such as the number of TV channels supported by the network, the average throughput and achieved quality for users in the network.

1.5 Outline of the Thesis

This remainder of the thesis is organized as follows. Chapter 2 presents some background knowledge and literature survey on scheduling in wireless networks. Chapter 3 describes the system model and research topics addressed in this thesis. The scheduling for nrtPS is discussed in Chapter 4, where a flexible resource allocation and scheduling framework is proposed and analyzed by jointly considering AMC technique and ARQ mechanisms. Chapter 5 addresses the scheduling for BE service, where WPF is proposed for achieving an efficient and flexible resource allocation among different SSs considering the heterogenous feature. Furthermore, an analytical model is developed to investigate the performance of WPF and evaluate the impact of the weights of SSs and channel conditions on some important performance metrics, such as service probability, spectral efficiency, throughput, and fairness. The scheduling for multicast transmission is studied in Chapter 6, where a cooperative multicast scheduling scheme is proposed to achieve a good fairness and high throughput for supporting multimedia services. Finally, concluding remarks and discussions on further work are given in Chapter 7.

Chapter 2

Background and Literature Review

2.1 IEEE 802.16 Broadband Wireless Network

In the past few years, there has been an increasing interest shown in wireless techniques for providing broadband wireless access, as an alternative to wired CM or DSL services. Following this trend, the IEEE sets up the 802.16 working group on broadband wireless access standards in 1999. This working group is focused on the specification for the global deployment of broadband WMANs and has released a series of IEEE 802.16 standards which define the air interface between a BS and multiple SSs. In specific, the current IEEE 802.16 standards identify the structures of two lowest layers: the PHY layer and the MAC layer. The PHY layer takes care of the establishment of the physical connection between the BS and SSs. IEEE 802.16 standards consider the frequency band in the range of 2-66 GHz, which is divided into two parts: 1) the range from 2 to 11 GHz is designed for non-line-of-sight (NLOS) transmissions; 2) the range between 10-66 GHz is designed for line-of-sight (LOS) transmissions. The MAC layer is composed of three sub-layers: the convergency sub-layer (CS), the common part sub-layer (CPS), and the security sub-layer. The main function of CS is to classify and map the data units received from the upper layer into appropriate connection identifiers (CID), and deliver/receive CS date units to/from the lower sub-layer CPS. CPS is an important part of the MAC layer, which is responsible for: 1)connection establishment and maintenance; 2) resource management and scheduling; 3) bandwidth requests and allocations; and 4) frame construction. Security sub-layer deals with security issues, such as packet encryption and key management, which are out of the scope of this thesis.

As a promising broadband wireless access standard, IEEE 802.16 has attracted extensive attentions from both the industry and academia [26–31]. The features of easy deployment, wide coverage, and high data rate make the IEEE 802.16 broadband wireless technique a prospective alternative to DSL and CM to provide broadband wireless access services in metropolitan area networks. Many leading telecommunications equipment manufacturers and service providers, such as Intel, Motorola, Fujitsu, Nokia, AT&T, etc., have shown great interest and support for the IEEE 802.16 broadband wireless access technique. They are members of the worldwide interoperability for microwave access (WiMAX)[1] Forum, which is an organization formed in 2001 to certify and promote the compatibility and interoperability of broadband wireless products based on the IEEE 802.16 standard. Two modes are supported in the IEEE 802.16 standard: mesh mode and Point-to-MultiPoint (PMP) mode. With the mesh mode, SSs can communicate with each other directly, extending the coverage of an IEEE 802.16 network. With the PMP mode, each SS directly communicates with the corresponding BS through wireless links, and the BS is connected to a core network through a wired link. Compared with the mesh mode, the PMP mode is more simple and easier to deployment due to its centralized control, and is expected to serve as an important role in the wireless metropolitan area network.

The first IEEE 802.16 standard was approved in December 2001. It is designed for point to point broadband wireless transmission with the frequency band of 10 to

[1] WiMAX is a telecommunications technology aimed at providing wireless access based on the IEEE 802.16 standard.

66 GHz. It uses a single carrier physical technique only with the capability of LOS transmission. As a further amendment to the IEEE 802.16 standard, IEEE 802.16a was ratified in January 2003, which aims to provide last mile fixed broadband wireless access with a data rate of up to 75 Mbps and a coverage of up to 30 miles. The frequency band is in the range of 2 - 11 GHz with the capability of NLOS transmission. The PHY layer is extended to include orthogonal frequency division multiplex (OFDM) and orthogonal frequency division multiple access (OFDMA) techniques. the IEEE 802.16-2004 was released in 2004, which takes the place of the earlier IEEE 802.16 documents, including IEEE 802.16a/b/c amendments. The IEEE 802.16e, an amendment to the IEEE 802.16-2004, was released in 2005, where the main enhancement is the support for mobility.

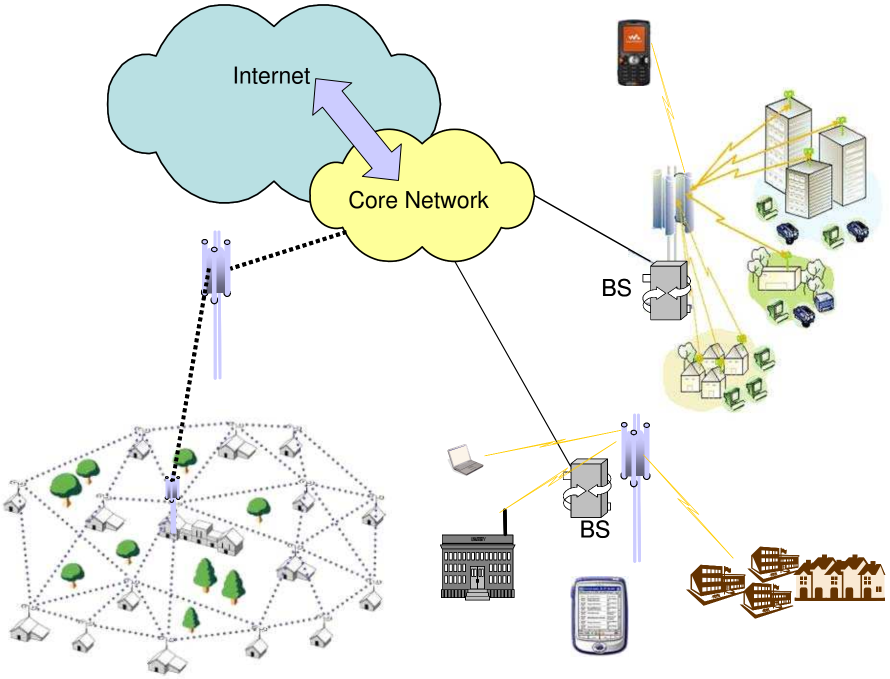


Figure 2.1: The Illustration of an IEEE 802.16 Broadband Wireless Megalopolitan Area Network.

Figure 2.1 illustrates an IEEE 802.16 broadband WMAN, which is composed of several IEEE 802.16 access networks and an IP-based core network. Multiple BSs are connected to the IP-based core network for efficiently managing and controlling the whole system. An IEEE 802.16 access network could work with mesh mode or PMP mode. There are four main business modes for IEEE 802.16 access networks: (1)they could be deployed in the rural areas for providing wireless broadband wireless services; (2)they could be deployed in the urban areas as an alternative or complimentary to DSL and CM wired access; (3)they could serve as the backhaul for the data collected from the WLAN access points; and (4)they could provide broadband data services for mobile users as well.

2.2 QoS Scheduling

An IEEE 802.16 network is designed to support heterogeneous types of services with different QoS requirements. QoS provisioning is a complicated issue which relies on an collaborative effort from different aspects, such as congestion control, CAC, and scheduling. Packet scheduling plays a particularly key role, which decides the order of packet transmissions, thus ensures that packets from different applications meet their QoS requirements. In general, main concerns on the design of a packet scheduling scheme are throughput, fairness, and QoS requirements. In IEEE 802.16 networks, SSs with good channel conditions can support a higher transmission rate than those with bad channel conditions. Therefore, assigning the network resource to SSs with good channel conditions can improve the network throughput and resource utilization. However, such a greedy approach may lead to a serious fairness problem, especially in IEEE 802.16 networks that aim to create a heterogeneous environment with QoS support. Thus, it is a critical issue to design efficient scheduling schemes by jointly considering throughput, resource utilization, fairness, and heterogenous QoS requirements.

2.2.1 Different Service Types

To fulfill service differentiation and QoS provisioning, the IEEE 802.16-2004 standard defines four types of services [32, 33]: Unsolicited Grant Service (UGS), rtPS, nrtPS, and BE service. UGS aims to support real-time constant-bit-rate (CBR) applications, such as T1/E1 classical pulse coded modulation (PCM) phone signal transmission and Voice over Internet Protocol (VoIP) without silence suppression, which are subject to a stringent delay and delay jitter constraints. rtPS is designed to support variable-bit-rate applications, such as Internet Protocol TV, gaming, and video conferences, where delay, minimum throughput, and maximum sustained throughput are defined and constrained. nrtPS is to support delay-tolerant applications, such as file transfer protocol (FTP), where minimum throughput is defined. BE traffic is subject to no QoS requirement. The extended real-time polling service (ertPS), the fifth type of service, is added in the IEEE 802.16e amendment. It is designed to support variable-bit-rate real-time applications, such as voice over IP with silent suppression.

2.2.2 Unicast Scheduling

Communications can be classified into three categories: unicast, multicast, and broadcast. Unicast is one-to-one communications, which denotes the sending of information packets to a single destination. Broadcast is the extreme opposite of unicast. It is the sending of information to all users in a network, while multicast is one-to-many communications, which denotes the delivery of data stream to multiple destinations simultaneously. In these three scenarios, scheduling deals with different issues due to their own unique features.

For unicast scheduling, many resource allocation and scheduling schemes have been developed for achieving a high throughput and good fairness by considering the wireless channel conditions [34]. Wireless fair scheduling is studied in [35], where short-term and long-term fairness and throughput are investigated. Channel-condition independent packet fair queueing is proposed in [36] to perform fair scheduling with guaranteed throughput. The scheduling scheme in [37] can achieve flexible scheduling and handle variable-size packets by combining the deficit round robin scheduling and an explicit compensation counter. However, the wireless channels in the aforementioned studies are modeled as either *good* or *bad* states. A user experiences error-free transmissions in a *good* channel state while unsuccessful transmissions in a *bad* channel state. Though simple, this channel model can hardly characterize a realistic wireless channel, where the achievable channel capacity varies by a smaller granularity.

With a more realistic wireless channel model, many studies are reported to deal with different types of services based on their intrinsic characteristics and QoS requirements. For UGS applications, a common solution is to periodically grant a fixed amount of resources since the service is designed for the CBR applications, and periodic grant can eliminate the overhead and latency caused by the transmission of bandwidth requests. Strict priority (SP) scheme is introduced in [38], where the UGS queue is assigned with the highest priority, followed by the rtPS, nrPS, and BE queues, respectively. Although class of services can be achieved, SP is obviously subject to a fairness problem. The scheme in [39] grants a relatively higher priority to the queues with stringent delay and throughput constraints, in which the system reserves relatively more resources to these queues.

For real-time applications, the main concern is to deal with the stringent delay requirement. The largest weighted delay first (LWDF) in [40] considers the head-of-line packet delay of each real-time queue. Two modified LWDF schemes are proposed in [19, 41], where a queue is selected for service so as to maximize the term $\gamma_j W_i r_j$, where γ_j is an arbitrary positive constant, W_i is the head-of-line packet delay, and r_j is the channel capacity for the queue j. The study in [42] defines a delay satisfaction indicator and a throughput satisfaction indicator in the design of the preference metrics for real-time and non real-time queues, respectively. In [43], delay information is considered in the scheduling. By monitoring the deadline violation of all queues, the proposed scheme can provide a good fairness and satisfy the delay requirements.

Since BE applications are not subject to any QoS requirement, the studies on the resource allocation and scheduling for the BE service have mainly focused on achieving high throughput and good fairness. Opportunistic scheduling (OS) [44] [45] aims at maximizing the system throughput by taking the best advantage of multi-user channel diversity, in which the resource is assigned to the queue with the best channel condition at a given time slot. An analytical model is developed in [46] to investigate the performance of OS in terms of the achieved throughput of each user. However, OS may result in poor fairness due to the starvation of end users persistently experiencing poor channel conditions for a long time. Many variants of OS are proposed [47–50]. Mobility information of users is considered in [50] to improve the system throughput. An opportunistic fair scheduling scheme is proposed in [20] to maintain the long-term fairness and achieve a high system throughput. Proportional fairness (PF) scheduling [22] [51] is proposed to initiate a compromise between the throughput and fairness among different users, where the long-term fraction of overall system resources obtained by each user is almost identical. PF can achieve a good fairness performance due to the identical long-term resource allocation. In addition, it takes advantage of multi-user channel diversity to obtain a high system throughput. However, it is generally difficult to conduct a quantitative analysis. A modified proportional fairness scheduling scheme is proposed in [52], where the scheduler selects a user with the highest ratio of the instantaneous channel condition to its average channel condition. By replacing the achieved average throughput with the average channel condition, the scheme is more tractable than the original proportional fairness scheme. The opportunistic fair scheduling, αPFS, is proposed in [53] to provide different fairness requirements by manipulating the parameter α. When $\alpha \to \infty$, max-min fairness is achieved, while proportional fairness is a special case when α=1. However, the αPFS scheme fails to operate in IEEE 802.16 networks since it cannot manipulate the achieved throughput of each SS. In [54], the selective relative best scheduling is proposed to schedule BE traffic, which aims to initiate a proper tradeoff between the fairness and system throughput by integrating the opportunistic scheduling and the relative best scheduling proposed in [55]. The scheme proposed in [56] takes the exponential rule for the delay-constrained traffic and the proportional fair scheduling for the BE traffic. A credit-based code division generalized process sharing scheme is proposed in [57], where the scheduler allocates resources based on both the generalized process sharing discipline and each user's credit to achieve high resource utilization as well as the long-term fairness. A cumulative distribution function based scheduling scheme is proposed in [58], where the probability of selecting a user for service depends on the distribution of its own channel conditions such that the fairness performance can be improved.

2.2.3 Multicast Scheduling

Multicast communications is an efficient approach for supporting one-to-many applications over a broadcast wireless channel and offers great opportunity for service providers to deliver TV, film, and other information (e.g., emergency alerts, software installation) to multiple users simultaneously. In recent years, wireless multicast services have attracted extensive attentions from both academia and industry. The Multimedia Broadcast Multicast Service (MBMS) has been standardized in various groups of the third generation partnership project (3GPP) [59] and is currently under active investigation. Since scheduling plays a key role in improving the wireless resource utilization and providing QoS for multicast multimedia services, this thesis will address efforts in the design of an efficient scheduling scheme for supporting multicast services.

In a multicast network, users requesting the same data can be logically grouped as an MGroup. For instance, all subscribers watching the same TV channel form an MGroup, and the total number of MGroups in the newtork equals the number of TV channels. Since mobile users are distributed at various locations and experience different fading and shadowing effects due to time-variant wireless channels, it is a very challenging issue to provide satisfying multicast services for all subscribers under a set of given design requirements. In general, fairness, throughput and reliability are three main concerns for designing an efficient multicast scheduling scheme. In [60], MGroups are served in a round-robin fashion with a fixed rate supported by the user at the edge of the cell. This scheme provides reliable multicast transmission at the expense of satisfying the high capability of users with good channel conditions. A multicast scheduling scheme is proposed for cellular systems, where MGroups are served in a round-robin fashion with a pre-defined transmission rate. For instance, the CDMA 2000 1xEV-DO networks use the fixed data rate of 204.8 Kbps for downlink multicast transmissions. Another approach is to select the the transmission rate supported by all group members, i.e., all group members can successfully decode the data at this rate. Thus, the group member with the worst channel condition becomes the bottleneck and the scheme results in a conservative resource utilization. This approach is especially inefficient when most users are in good channel conditions and capable for high rate transmissions while only a small fraction of users are too far away from the BS or suffer deep fading. These two multicast scheduling schemes underutilize wireless resources because they use conservative transmission rates to assure reliable multcast transmissions, without taking advantage of diverse channel conditions of multiple group members. To improve the network resource utilization, one possible approach is to split an MGroup into several subgroups which can support different rates. In [61], a scheme has been proposed to divide a cell into two service regions. The BS transmits two data streams with different power levels such that the users near the BS can successfully receive both of them while the users away from the BS only receive one data stream. This scheme can achieve a higher throughput than that in [60] due to the consideration of different channel conditions. However, it does not give details on how to efficiently select MGroups and how to guarantee the reliable transmission of users far from the BS.

A number of scheduling schemes have been proposed for achieving high throughput and good fairness performance. A proportional fair scheduling scheme is proposed in [62], where an MGroup and its corresponding transmission rate are dynamically selected based on the proportional fair policy, rather than the worst channel condition of group members. In [63], the inter-group proportional fairness scheme is proposed, where the BS selects MGroup and the transmission rate in such a way that the summation of $log(T_k^g)$ for all MGroups is maximized, where T_k^g is the group throughput for MGroup k. These two schemes achieve a good tradeoff between the throughput and the fairness, but they do not consider how to efficiently improve the throughput of the uses with bad channel conditions. Meanwhile, it is difficult to conduct a quantitative analysis since the selection of MGroups depends on the average throughput of each group member in each MGroup. In [64], a threshold based mutlicast scheme is proposed, in which the sender transmits only when a sufficient number of group members can successfully receive the data. In [65], the relation between the stability and throughput based on the threshold multicast scheduling is studied, which indicates the proposed scheme in [64] may lead to an unstable system when the threshold is not set properly.

2.3 Discussion and Summary

Packet scheduling plays an important role in fulfilling service differentiation and QoS provisioning. It deals with the resource allocation and multiplexing at the packet level. Thus, it is critical in improving the network resource utilization and providing different QoS among heterogeneous types of traffic. For unicast scheduling, many previous studies are focused on the real-time applications by paying more attentions on the stringent delay requirements and the BE service by considering the improvement of throughput and fairness performance. To the best of our knowledge, very few research efforts have delicately dealt with the resource allocation and scheduling for nrtPS flows although it accounts for a majority of network traffic. Although nrtPS flows are not delay-sensitive compared to rtPS, it is not acceptable if nrtPS flows are starved for a long time. Therefore, how to design an efficient scheduling scheme to flexibly control the experienced delay while achieving a high resource utilization and maintaining the minimum bandwidth requirements is a critical issue. In addition, most previous work on the scheduling for BE service is based on the homogeneity among end users. They can not work efficiently in IEEE 802.16 networks where SSs could be office buildings, residence houses, and mobile users. The heterogeneity of SSs poses new challenges in the design of scheduling schemes for achieving efficient and fair resource allocation based on the traffic demands/patterns of different SSs in IEEE 802.16 networks. For multicast scheduling, many previous studies focus on alleviating the negative impact caused by the diverse channel conditions of multiple members in an MGroup. However, they do not exploit any potential advantages provided by the channel diversity in multicast scenarios. For instance, since multiple group members are independently located in the network, a signal which is received weakly by one group member may be successfully received by other nearby group members. If the information received by group members in an MGroup could be shared within each other, more efficient and reliable multicast transmissions can be achieved, and the network throughput can be improved significantly.

Chapter 3

System Model

3.1 Network Model

We consider an IEEE 802.16 access network with PMP mode, as shown in Figure 3.1. It is composed of a BS and multiple SSs. The SSs directly communicate with the BS through wireless channels, while the BS connects to the core network through a wired connection. Generally, an SS could be a residence house, a mobile user, or an office building providing the Internet service to multiple customers. The SSs associated to the BS have different channel conditions, which depend on their geometric locations and the impact of the short-term fading.

3.2 Channel Model

3.2.1 Rayleigh Flat Fading Channel

Since radio signals are transmitted in an open space, they suffer from signal reflection, diffraction, and scattering. In this thesis, we consider both large-scale path-loss attenuation and small-scale fading in the channel model. Path-loss attenuation is mainly determined by the geographical environment and distance between the receiver and the

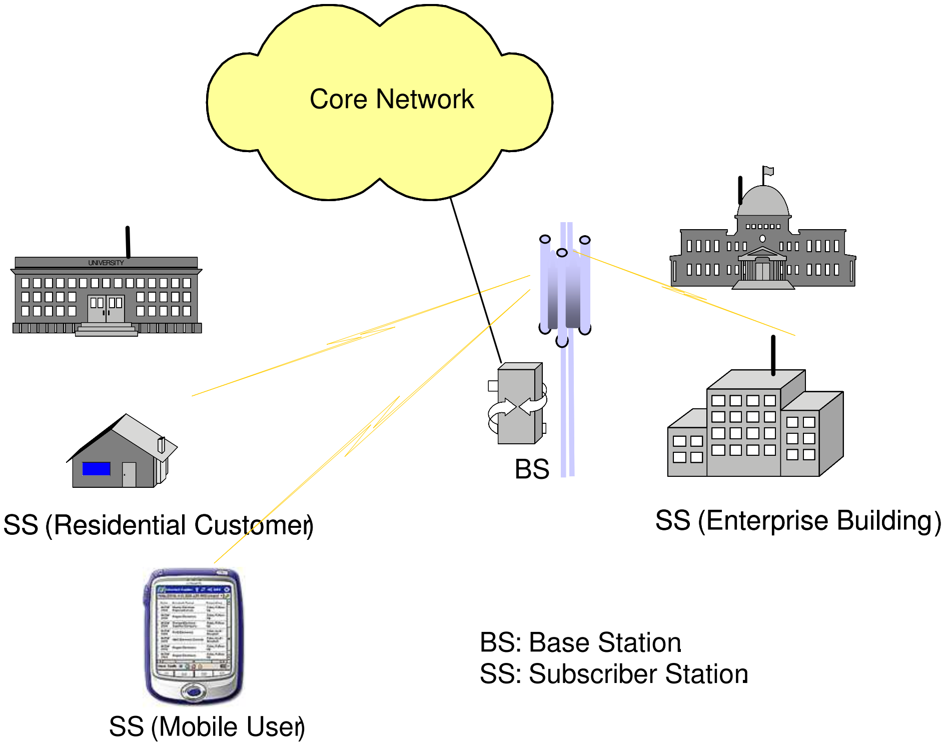


Figure 3.1: The illustration of an IEEE 802.16 Access Network.

transmitter, which can be modeled as (in decibels) [66]

$$PL(d) = PL(d_0) + 10\ k\ log_{10}(\frac{d}{d_0}) \tag{3.1}$$

where $PL(d_0)$ is the pass loss at the close-in reference distance d_0, k is the path loss exponent, and d is the distance between the receiver and the transmitter.

Small-scale fading is caused by multiple versions of a transmitted signal with different delay times such that it occurs spontaneously in the time span with a random duration and depth, and is also considered independent of the large scale path loss. Rayleigh flat fading channel, a common used channel model for NLOS transmissions, is adopted to describe the small-scale fading, where the perceived signal-to-noise ratio (SNR) of an SS at each MAC frame is a random variable with an exponential distribution and its probability density function (p.d.f.) is given by

$$f(\gamma) = \frac{1}{\overline{\gamma}} e^{-\frac{\gamma}{\overline{\gamma}}} \tag{3.2}$$

where γ and $\overline{\gamma}$ represent the instantaneous and average received SNR, respectively.

3.2.2 N-state Markov Channel Model

Rayleigh flat fading channel is a continuous state channel model. In some cases, it is necessary to model a wireless channel as a discrete process for performance analysis. Wireless channels suffer from deep fading that occurs randomly in the time span with a random duration and depth. Numerous studies have shown that such channels can be described by a Markov model to capture the bursty error nature. A simple Markov model is the two-state Gilbert-Elliot model, which is composed of two states: *Good* and *Bad*. The *Good* state is the one with a low error probability while the *Bad* state is the one with a high error probability [67–69]. The two-state model is too simple to represent the realistic channel conditions. To characterize the wireless channels more accurately, in this thesis, an N-state Markov channel model [70] is used to describe the time-varying channel conditions of each SS. The boundaries of SNR for the N-state Markov channel model is denoted as a row vector $B = [b_0, b_1, \cdots, b_N, b_{N+1}]$, where $b_0 = 0$ and $b_{N+1} = \infty$. When the received SNR is located in the set $[b_n, b_{n+1})$, the channel state is represented by the state n. In addition, when the channel model is constructed, the AMC technique, which has been defined in the IEEE 802.16 standard, is taken into account. AMC is an advanced technique at the PHY layer to achieve a high throughput by adaptively adjusting the sending rate according to the channel conditions. With AMC, the received SNR is divided into several disjoint regions. Based on the perceived SNR, the BS selects a proper modulation level and coding rate. According to the IEEE 802.16 standard, an 8-state Markov channel model is used in this thesis. The value of b_i and the modulation and coding level corresponding to each channel state are specified in Table 3.1.

Table 3.1: State boundaries and information bits carried by an OFDM symbol.

State ID	Modulation level and coding	Information bits/OFDM symbol (I_n)	b_n(dB)
0	*silent*	0	0
1	BPSK(1/2)	96	3
2	QPSK(1/2)	192	6
3	QPSK(3/4)	288	8.5
4	16QAM(1/2)	384	11.5
5	16QAM(3/4)	576	15
6	64QAM(2/3)	768	18.5
7	64QAM(3/4)	864	21

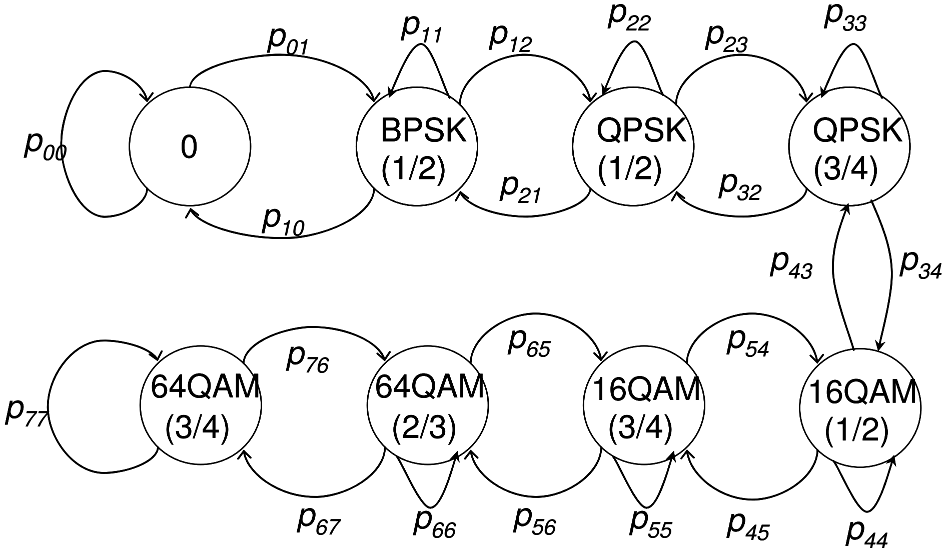


Figure 3.2: The finite state channel model.

The finite state Markov channel (FSMC) model can be abstracted as shown in Figure 3.2. State 0 represents the state with no transmission permitted, which happens when the channel condition is very poor. In this case, the corresponding queue should not transmit any data in order to improve the system throughput. For simplicity, states

$BPSK(1/2)$, $QPSK(1/2)$,$\cdots$, and $64QAM(3/4)$ are represented by symbols 1, 2,$\cdots$, and 7, respectively.

The probability of staying at state n (denoted as $\pi(n)$) is given by [70]

$$\pi(n) = \frac{\Gamma(m, m \cdot b_n/\bar{\gamma}) - \Gamma(m, m \cdot b_{n+1}/\bar{\gamma})}{\Gamma(m)} \tag{3.3}$$

where $\bar{\gamma}$ is the average SNR, m is Nakagami fading parameter, $\Gamma(m)$ is the Gamma function, and $\Gamma(m, \gamma)$ is the complementary incomplete Gamma function. Rayleigh fading channel is a special case when $m = 1$. For a slow fading channel (i.e., state transition occurs only between adjacent states), the state transition matrix for the FSMC can be expressed as follows

$$P = \begin{bmatrix} p_{00} & p_{01} & 0 & \cdots & 0 & 0 & 0 \\ p_{10} & p_{11} & p_{12} & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & p_{N-2,N-1} & p_{N-1,N-1} & p_{N-1,N} \\ 0 & 0 & 0 & \cdots & 0 & p_{N,N-1} & p_{N,N} \end{bmatrix} \tag{3.4}$$

The transition probability from the state n to k (denoted as p_{nk}) is obtained as follows

$$p_{n,n+1} = \frac{L_{n+1} \cdot T_M}{\pi(n)} \quad n = 0, 1, \cdots, N-1 \tag{3.5}$$

$$p_{n,n-1} = \frac{L_n \cdot T_M}{\pi(n)} \quad n = 1, 2, \cdots N \tag{3.6}$$

$$p_{n,n} = \begin{cases} 1 - p_{n,n+1} - p_{n,n-1}, & 0 < n < N \\ 1 - p_{0,1}, & n = 0 \\ 1 - p_{N,N-1}, & n = N \end{cases} \tag{3.7}$$

where T_M is the time duration of a MAC frame, and L_n is the level crossing rate at b_n corresponding to the state n, which can be estimated by

$$L_n = \sqrt{2\pi \frac{m \cdot b_n}{\bar{\gamma}}} \cdot \frac{f_d}{\Gamma(m)} \cdot (\frac{m \cdot b_n}{\bar{\gamma}})^{m-1} \cdot exp(\frac{m \cdot b_n}{\bar{\gamma}}) \tag{3.8}$$

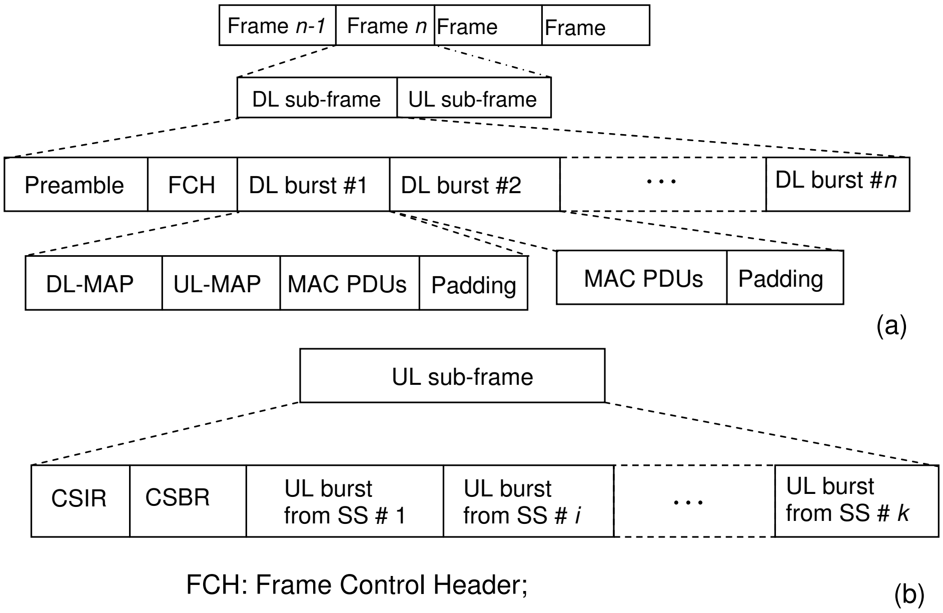
where f_d is the Doppler frequency.

3.3 Media Access Control Structure

The communication paths between the BS and SSs are divided into uplink (UL) channel (from SSs to the BS) and downlink (DL) channel (from the BS to SSs). In an IEEE 802.16 network, the uplink channel is shared by all the SSs based on the time division multiple access (TDMA) mechanism, while the BS broadcasts information based on the time division multiplexing (TDM) mechanism over the downlink channel. Two UL/DL duplexing alternatives are specified in IEEE 802.16 standard: one is time division duplexing (TDD) and the other is frequency division duplexing (FDD). In this thesis, we focus on the TDD-OFDM[1] /TDM downlink scheduling.

Figure 3.3 gives a sketch of the MAC structure for an IEEE 802.16 network. At the MAC layer, the time domain is divided into MAC frames with equal durations. A MAC frame is composed of a downlink sub-frame (DL sub-frame), followed by an uplink sub-frame (UL sub-frame). A DL sub-frame consists of a preamble signal, which is used for synchronizing SSs with the BS, followed by frame control header (FCH), downlink MAP (DL-MAP), uplink MAP (UL-MAP) messages, and several downlink transmission bursts. DL-MAP and UL-MAP messages specify the allocation of the transmission bursts among SSs, including the corresponding time duration and burst profiles such as the modulation level and coding rate. A burst is a contiguous portion of data stream using the same physical parameters, such as modulation level and preamble length. An UL sub-frame is composed of the contention slot for initial ranging (CSIR), the contention slot for bandwidth request (CSBR) and several UL bursts from different SSs. CSIR is used for SSs to join the network. When an SS attempts to join the network, it sends initial ranging messages during this slot to acquire the operation

[1] OFDM (Orthogonal Frequency Division Multiplexing) is an advanced communication techniques adopted in IEEE 802.16 networks to address frequency-selective fading due to multi-path propagation.



FCH: Frame Control Header;
CSIR: Contention Slot for Initial Ranging;
CSBR: Contention Slot for Bandwidth Request.

Figure 3.3: The MAC Structure for an IEEE 802.16 Network.

parameters, such as frequency, transmission power level. CSBR is used for SSs to send their bandwidth request messages. In this thesis, we focus on the downlink scheduling.

3.4 Automatic Repeat reQuest Mechanism

Automatic Repeat reQuest (ARQ), a link-layer close-loop error control mechanism, has been identified as an efficient approach to achieve a low packet loss rate in an error-prone wireless environment. ARQ with cumulative acknowledgement (referred to as the cumulative ARQ), ARQ with selective repeat acknowledgement (referred to as the selective ARQ), and the hybrid ARQ are three main ARQ mechanisms supported by IEEE 802.16 networks. The cumulative ARQ is simple, but less efficient than the selective ARQ and the hybrid ARQ. The hybrid ARQ is a variation of the ordinary

ARQ mechanisms (*i.e.* the cumulative ARQ and the selective ARQ) by combining with forward error correction (FEC). It is efficient in poor channel conditions with the expense of more complex decoding procedure and more memory requirements. The selective ARQ is more efficient than the cumulative ARQ, and more simple and easy to implement than the hybrid ARQ. In this thesis, we focus on the cumulative ARQ and selective ARQ, which are jointly considered in the proposed scheduling scheme for nrtPS applications. These two ARQ mechanisms will be elaborated further as follows.

Without loss of generality, the following discussions are for downlink transmissions. The principle of the cumulative ARQ in IEEE 802.16 networks is described as follows. At the BS, the link-layer entity receives service data units (SDUs) from the upper layer. The received SDUs are classified and buffered into the corresponding queues based on their associated SSs and their service types. Each SDU is segmented into several protocol data units (PDUs) of equal size prior to delivery over the wireless channel. Time domain is divided into MAC frames with a common duration. At the beginning of each MAC frame, the BS assigns timeslots for transmitting PDUs of each SS according to the adopted scheduling scheme. If a queue obtains the chance of transmissions, a certain number of PDUs buffered at this queue are transmitted at this DL sub-frame. Otherwise, no transmission is permitted for this queue during this DL sub-frame. At the receiver side, the received PDUs are buffered until a complete SDU is successfully received. Each SS sends feedback information back to the BS during each UL sub-frame, indicating whether PDUs are successfully received or not. One of the important feedback information is the fragment sequence number (FSN), which indicates the sequence number of the last received PDU before the first lost PDU. Based on this feedback information, the BS has the knowledge of the sequence number of the first lost PDU among all PDUs launched in this DL sub-frame, which is (FSN + 1). Then the BS updates the sequence number of its sending window as (FSN + 1). All PDUs launched in this DL sub-frame with a sequence number larger than or equal to (FSN + 1) have to be retransmitted no matter they have been successfully received or not. When a scheduling scheme is taken into account, these PDUs are retransmitted when the corresponding queue obtains the transmission opportunity, instead of being retransmitted immediately in the next MAC frame.

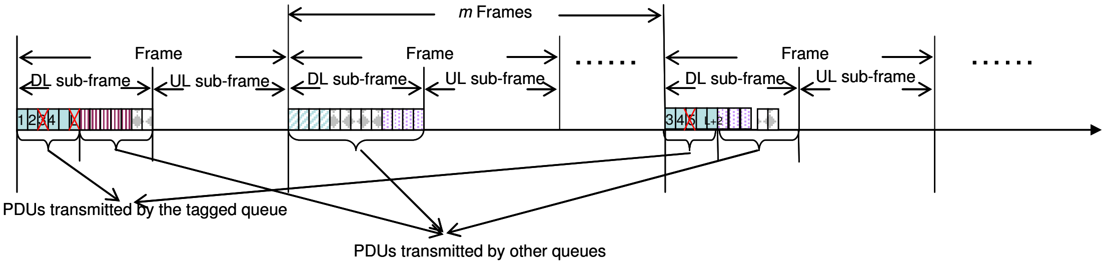


Figure 3.4: The illustration of the cumulative ARQ mechanism.

Figure 3.4 illustrates the cumulative ARQ mechanism with a scheduling scheme for a tagged queue. In the first frame, since the PDU with the sequence number 3 has been lost and identified by the receiver, the largest sequence number among all continuously successfully received PDUs is 2. Therefore, in the feedback information, the FSN is 2, and all the PDUs with a sequence number from 3 to L need to be retransmitted. Then, the tagged queue waits for another m frames until it obtains the transmission opportunity again, where m is a random variable depending on the deployed scheduling scheme. When the tagged queue obtains the transmission opportunity, L PDUs with sequence number from 3 to L+2 are retransmitted/transmitted. Then, the BS updates the FSN accordingly based on the feedback information.

Another ARQ mechanism is the selective ARQ, where the receiver keeps track of the sequence number of received PDUs and send back the ACK/NACK to the BS at the following UL sub-frame to report the information about whether or not the PDUs transmitted at the current DL sub-frame are successfully received or not. Based on the feedback information, only failed PDUs are retransmitted next time when this queue obtains the transmission opportunity. Since less retransmissions are involved, the selective ARQ is more efficient than the cumulative ARQ. On the other hand, due to out-of-order sequence numbers, the implementation of the selective ARQ need more buffer and complexity compared with the cumulative ARQ.

3.5 Research Topics

Scheduling is a critical issue in providing QoS satisfaction among heterogenous types of services for any kind of networks. IEEE 802.16 network is no exception. This thesis is to contribute toward efficient and flexible scheduling for different types of services in IEEE 802.16 networks to achieve a satisfying performance in terms of throughput, resource utilization, fairness, and QoS provisioning.

As discussed in Section 2.2, many previous work on the scheduling od focused on the rtPS and BE service. However, to the best of our knowledge, very few research efforts have delicately dealt with the resource allocation and scheduling for nrtPS in spite of its ultimate importance. It has been reported that rtPS flows accounts for a majority of the Internet traffic [71, 72]. With such predominant bandwidth consumption, it is crucial to develop a dedicated strategy for dealing with nrtPS traffic in IEEE 802.16 networks. In addition, ARQ is a widely-used close-loop error control mechanism to achieve a low packet loss rate in error-prone wireless environments. It should be jointly considered in the scheduling scheme to evaluate the achieved goodput at the receiver side, rather than the assigned resource at the transmitter side. Although ARQ mechanisms have been extensively studied [73–77], some new challenges have emerged with the launching of the IEEE 802.16 standard. The IEEE 802.16 standard specifies some advances in physical layer techniques and media access control protocol, such as AMC and the flexible retransmission of lost packets. These characteristics have posed fundamental differences in the efforts of performance analysis of ARQ in IEEE 802.16 networks compared with that in many previous works. Firstly, most of previous studies simply assume that the time taken to transmit a packet is a constant, generally defined as a timeslot. Such an assumption does not hold in IEEE 802.16 networks due to the adoption of AMC. With AMC, the link capacity varies with the channel conditions, leading to the fact that the time taken to retransmit a packet varies with the current channel condition, and may be different from that taken in the previous transmission attempt. Secondly, most of the former studies commonly assume that the lost packets are retransmitted immediately in the very next timeslot [74, 76, 78, 79]. When an efficient scheduling scheme is considered, the retransmission of lost packets is flexible and does not just occur within the very next timeslot. For instance, when the channel condition of the next timeslot is poor, the retransmission could experience a much higher packet loss rate. Therefore, a fundamental difference can be identified in the design and performance analysis in IEEE 802.16 networks compared with that in previous studies. Thus, it is critical to develop an efficient resource allocation and scheduling framework that can not only maintain the minimum bandwidth requirements for nrtPS flows by jointly considering ARQ mechanisms at the MAC layer and AMC at the PHY layer, but also initiate a graceful compromise between the resource utilization and delivery delay.

On the other hand, many previous studies on the scheduling for BE service mainly investigate how to evenly allocate the available resource among users. The fairness in these scheduling schemes is defined on the basis that the traffic load and allocated resource are identical/homegeneous among all users. However, the requirement of homogeneous users is less likely to be satisfied in IEEE 802.16 networks. As shown in Figure 3.1, an SS in an IEEE 802.16 network could be a residential house, a mobile user, or an office building providing the Internet service to many customers. Multiple types of SSs is one of the unique features for IEEE 802.16 networks, compared with the conventional wireless communication systems such as the 3G cellular systems where all end users are individual cell phones or handsets. Due to the multiple types of SSs, each SS may submit a much different long-term traffic demand patterns. For instance, BE traffic load for an SS of office building could be much higher than that for an SS of resident house during the day time, while residence houses could be the major bandwidth consumers during the late evening. Therefore, an efficient scheduling scheme for BE service in IEEE 802.16 networks should be able to allocate the available resources among different SSs in an adaptive and flexible way, such that the network operators can freely perform the bandwidth allocation among SSs according to their different traffic load/demand patterns.

Beside the aforementioned unicast scheduling, multicast communications enables efficient one-to-many transmissions over a broadcast wireless channel for supporting multicast applications, such as IPTV, mobile TV, emergency alerts, software installation, which have attracted great attentions from both industry and academia. They are expected to serve as killer applications in the next generation IEEE 802.16 WMANs and will contribute immense market value to the service providers. Therefore, it is critical to design an efficient scheduling scheme for supporting broadband multimedia services over IEEE 802.16 networks. In general, users demanding the same copy of data are logically grouped as an MGroup, and each user is a group member. In a multicast scenario, group members are distributed at different geometric locations with diverse channel conditions. The diverse channel conditions among group members of an MGroup is the main challenge for designing multicast scheduling. First, due to the diverse channel conditions of different group members of an MGroup, ARQ is not efficient for recovering the lost packets in the multicast scenario. The retransmission of the packet lost at one group member may lead to a waste of bandwidth for the group members in good channel conditions. As discussed in Section 2.2.3, many previous studies focus on improving the group throughput at the expense of the reliability of the group members with bad channel conditions. However, in most cases, it is necessary to provide satisfying services to all users in multimedia multicast applications, no matter when the channel condition is good or bad. Therefore, it is critical to design an efficient multicast scheduling scheme for not only providing high throughput for the group members with good channel conditions, but also improving the throughput for group members with bad channel conditions.

3.6 Summary

In this chapter, we have described the network model of a general IEEE 802.16 wireless access network and introduced the MAC structure. ARQ mechanisms have also been briefly introduced, which will be considered in the proposed scheduling scheme discussed in the following Chapters. Finally, we have discussed the research topics on the scheduling for supporting heterogenous types of services with different QoS requirements.

Chapter 4

Scheduling for Non-real-time Polling Service

Packet scheduling in wireless networks have been extensively studied in the previous work. However, most of them focused on either BE applications by improving the system throughput and fairness performance, or real-time applications with the delay constraint being the main concern as discussed in Subsection 2.2.2.

nrtPS is an important kind of applications, which accounts for the majority of the Internet traffic. It has been reported that transmission control protocol (TCP) traffic (which exclusively takes nrtPS as the carrier in IEEE 802.16 networks) may take up to 80% of the total Internet bandwidth [80]. With such predominant bandwidth consumption, it is crucial to develop a dedicated strategy for dealing with nrtPS traffic in IEEE 802.16 networks. A few studies have considered non-real-time applications, but they only focused on satisfying either the throughput ratio or the acceptable delay [81–84].

An efficient resource allocation and scheduling for nrtPS traffic is to satisfy the minimum throughput requirement as well as achieving a high resource utilization with acceptable delay. Efforts on improving resource utilization and reducing experienced delay are in general contradictory since high resource utilization can be achieved by assigning the resource to the SSs with the best channel conditions, while leaving the SSs with poor channel conditions starved and experienced a long delay. Although nrtPS connections are delay-tolerant, they should not be starved for too long since otherwise the flows will suffer considerable performance degradation. For instance, based on the 3GPP standard, the delay for the low delay data service and the long constrained delay data service should be less than 50ms and 300ms, respectively [85]. Therefore, how to compromise the resource utilization and experienced delay of each SS is a challenging yet important issue.

4.1 A Flexible Resource Allocation and Scheduling Framework for Non-real-time Polling Service

In this section, we propose a simple yet efficient resource allocation and packet scheduling framework for nrtPS traffic in IEEE 802.16 networks such that the minimum bandwidth requirements can be satisfied while a flexible tradeoff between the packet delivery delay and the resource utilization is initiated. The flowchart of the proposed framework is shown in Figure 4.1. The parameters h and L are two key operation parameters in the proposed framework. h is the number of SSs selected at each MAC frame, which is the parameter controlling the tradeoff between the resource utilization and the packet delivery delay of each SS, while L (in the unit of PDUs) is the bandwidth granted to an SS when it is being served, which is the parameter to fulfill the satisfaction of the minimum bandwidth requirement of an nrtPS flow. By manipulating these two parameters, the framework can satisfy the minimum bandwidth requirement of each nrtPS flow and fulfill the flexible tradeoff between the delivery delay and the resource utilization. Furthermore, the proposed framework takes the channel condition of each SS into account. By exploiting the multi-user channel diversity, the proposed scheme can achieve a high system throughput.

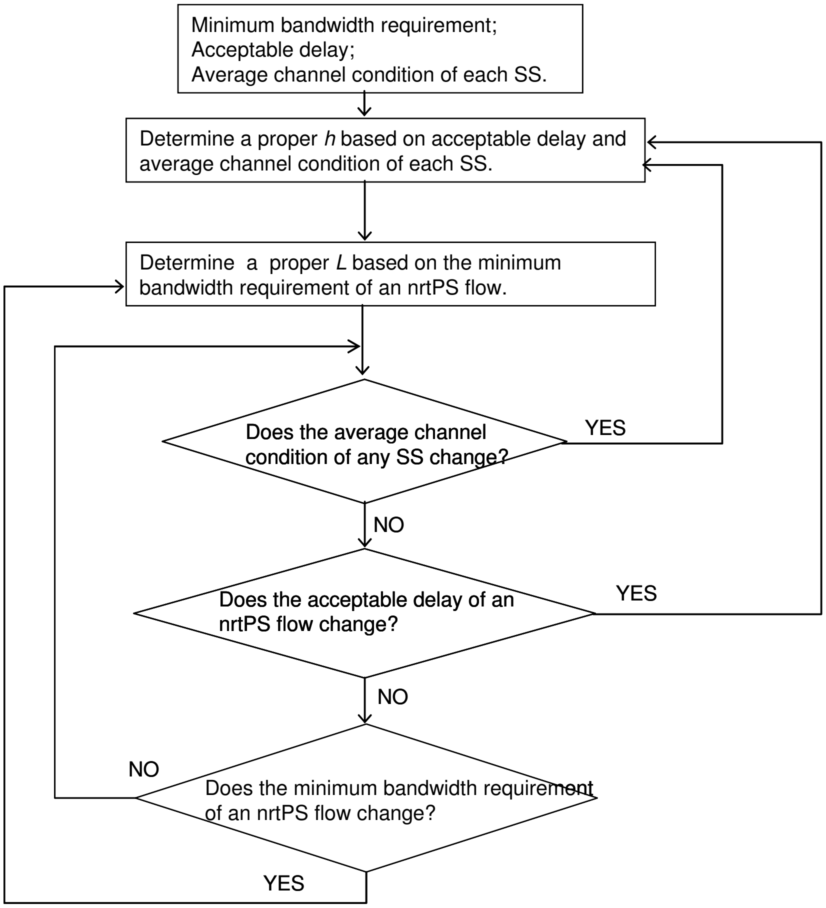

Minimum bandwidth requirement;
Acceptable delay;
Average channel condition of each SS.
Determine a proper h based on acceptable delay and average channel condition of each SS.
Determine a proper L based on the minimum bandwidth requirement of an nrtPS flow.
Does the average channel condition of any SS change?
YES
NO
Does the acceptable delay of an nrtPS flow change?
YES
NO
NO
Does the minimum bandwidth requirement of an nrtPS flow change?
YES

Figure 4.1: The flowchart of the proposed resource allocation and scheduling framework.

Due to the error-prone wireless channel, there may exist much gap between the assigned bandwidth and the achieved goodput. The proposed framework takes this situation into account by jointly considering ARQ mechanisms at the MAC layer. With the use of ARQ, the information about whether or not the PDUs transmitted at a DL sub-frame are successfully received is sent back to the BS at the following UL sub-frame. Based on the feedback information, the failed PDUs are retransmitted next time when this queue obtains the transmission opportunity, instead of being retransmitted immediately. Two ARQ mechanisms (*i.e.* the cumulative ARQ and the selective ARQ discussed in Section 3.4) are considered in the proposed framework. By taking into account the impact of ARQ mechanisms on the retransmission of lost PDUs, we analyze the performance of the proposed framework in terms of the achieved goodput and the packet delivery delay for each SS. Furthermore, the proposed framework also considers the AMC technique at the PHY layer. With the adoption of AMC, the system dynamically adjusts the modulation level according to the channel condition of each SS. In IEEE 802.16 networks, different modulation levels lead to different numbers of information bits carried by an OFDM symbol. Therefore, resource utilization is one of the most important metrics to evaluate the performance of the proposed framework, which is also analyzed in this chapter.

Specifically, the proposed framework works as follows. At the beginning of each MAC frame, h SSs with better channel conditions are selected and granted with transmission opportunities at this DL sub-frame. Each selected SS is assigned with a specific amount of resources, which is denoted as L, according to the minimum throughput requirement of its nrtPS flow and the channel conditions of all SSs. By manipulating the parameters h and L properly, the proposed resource allocation and scheduling framework can fulfill any possible design requirement, such as resource utilization, throughput requirements and delivery delay requirements. When h is set to 1, the scheduling framework is degraded to the opportunistic scheduling, which can obtain the maximum resource utilization at the expense of possibly long delivery delay of the SSs with poor channel conditions. When h equals to the total number of SSs associated to the BS, the resource utilization is low, but the delivery delays of SSs are small. Meanwhile, the setting of parameter L depends not only on the minimum bandwidth requirement of a nrtPS flow, but also on the channel conditions of all SSs associated to the BS.

4.2 Performance Analysis

In order to study the performance of the proposed framework and the impact of the two parameters on the design requirements, such as resource utilization, minimum bandwidth requirement, and packet delivery delay, an analytical model is developed to evaluate some key performance metrics, including inter-service time, PDU delivery delay, SDU delivery delay, goodput, and resource utilization. For simplicity, the SS under consideration is referred to as the tagged SS; and at the BS, the queue that buffers the nrtPS PDUs associated to the tagged SS is referred to as the tagged queue. When the tagged queue obtains the transmission opportunity, L PDUs buffered at this queue are transmitted at this DL sub-frame. The following assumptions are made in the performance analysis:

(1) a link layer SDU corresponds to an IP packet;

(2) each SDU is fragmented to F PDUs with an equal size of B bits;

(3) feedback information of PDUs transmitted at a DL sub-frame is sent back to the BS at the following UL sub-frame using the UL-ACK channel, which has been defined in IEEE 802.16e standard;

(4) resources are available for nrtPS traffic admitted into the network at each MAC frame. This assumption is reasonable provided with a well-defined connection admission control strategy adopted in the network;

(5) when a queue is scheduled, it has PDUs waiting for transmission.

4.2.1 Performance Analysis with Cumulative ARQ

- **Service Probability of SS_i**

The service probability of SS_i is defined as the probability for SS_i to obtain the chance of service at an arbitrary DL sub-frame. Firstly, we classify all SSs into three groups based on the channel state of the tagged SS at a specific MAC frame. Given the channel state of the tagged SS is at state n, the three groups are composed of the SSs with channel conditions better than, same as, and worse than the state n , respectively, which is denoted as the groups G_1, G_2, and G_3, respectively. Let k_1, k_2 and k_3 denote the number of SSs in the groups G_1, G_2, and G_3, respectively. The tagged SS, which belongs to the group G_2, obtains the chance of transmission only when the condition $k_1 < h$ holds. Otherwise, all the selected SSs should come from the group G_1. When the condition is satisfied, the probability that the tagged SS obtains the chance of transmission can be derived based on the values of k_1 and k_2. Since the total number of selected SSs is h, and k_1 SSs are at the channel states better than state n, $h - k_1$ is the quota left for the SSs in the groups G_2 and G_3. When $h - k_1$, is larger than k_2, all the SSs in the group G_2 are selected, i.e., the tagged queue obtains the chance of transmission at this DL sub-frame with probability 1. On the other hand, when $k_2 > h - k_1$, the BS randomly selects $h - k_1$ out of k_2 SSs in the group G_2. Therefore, the tagged queue obtains the chance of transmission with a probability $(h - k_1)/k_2$. To take these situations into account, we define the function $\xi(\cdot)$, as given in (4.1). k_1 is a value in the set $[0, h-1]$ such that the tagged SS can obtain the chance of transmission. Let M be the total number of SSs in the network. Under a specific value of k_1, k_2 is a value in the set of $[1, M - k_1]$. The set begins with 1 since at least the tagged SS is in the group G_2. When the values of k_1 and k_2 are given, k_3 is $M - k_1 - k_2$.

$$\xi(\frac{h-k_1}{k_2}) = \begin{cases} 1, & h - k_1 \geq k_2 \\ \frac{h-k_1}{k_2}, & h - k_1 < k_2 \ . \end{cases} \tag{4.1}$$

Let $\Omega(j_1)$, $\Omega(j_2)$, and $\Omega(j_3)$ denote the sets of SSs in the groups G_1, G_2, and G_3, respectively. Thus, the probability that the tagged SS obtains the chance of service and stays at the state n along with the specific $\Omega(j_1)$, $\Omega(j_2)$, and $\Omega(j_3)$ can be expressed as

$$\xi(\frac{h-k_1}{k_2})\left[\prod_{i_1\in\Omega(j_1)} Pr(S_{i_1}>n)\prod_{i_2\in\Omega(j_2)} Pr(S_{i_2}=n)\prod_{i_3\in\Omega(j_3)} Pr(S_{i_3}<n)\right] \tag{4.2}$$

where the function $\xi(\cdot)$ is defined as (4.1), while $\prod_{i_1\in\Omega(j_1)} pr(S_{i_1}>n)$, $\prod_{i_2\in\Omega(j_2)} pr(S_{i_2}=n)$, and $\prod_{i_3\in\Omega(j_3)} pr(S_{i_3}<n)$ denote the probability that all SSs in groups G_1, G_2, and G_3 are at the channel states better than, same as, and worse than state n, respectively. For instance, the system is composed of SS_1, SS_2, SS_3, SS_4, SS_5, and SS_6. The tagged SS is SS_1, which stays at the channel state $n=3$. Let $h=3$, $k_1=2$, and $k_2=3$. The probability that SS_1 obtains the chance of transmission with the channel state 3 while $\Omega(j_1)$, $\Omega(j_2)$, and $\Omega(j_3)$ are $\{SS_2,\ SS_3\}$, $\{SS_1,\ SS_4,SS_5\}$, and $\{SS_6\}$, respectively, is given by

$$\frac{1}{3}\left[\prod_{i_1\in\{SS_2,SS_3\}} Pr(S_{i_1}>3)\prod_{i_2\in\{SS_1,SS_4,SS_5\}} Pr(S_{i_2}=3)\prod_{i_3\in\{SS_6\}} Pr(S_{i_3}<3)\right] \tag{4.3}$$

Equation (4.2) is for the specific $\Omega(j_1)$, $\Omega(j_2)$, and $\Omega(j_3)$. In the following, the number of all possible $\Omega(j_1)$, $\Omega(j_2)$, and $\Omega(j_3)$ are taken into consideration. Let a_1 represent the number of possible combinations for selecting k_1 SSs out of $(M-1)$ SSs to construct the group G_1, where M is the total number of SSs in the network. After SSs in the group G_1 are selected, there are $(M-k_1)$ SSs left. Let a_2 represent the number of possible combinations for selecting (k_2-1) SSs out of the left $(M-k_1-1)$ SSs to construct the group G_2. At last, the left $M-k_1-k_2$ SSs consist of the group G_3. We have $a_1=\binom{M-1}{k_1}$ and $a_2=\binom{M-k_1-1}{k_2-1}$. In other words, given k_1, the total number of possible $\Omega(j_1)$ is a_1, and the set of all possible $\Omega(j_1)$ is represented by $\{\Omega(j_1), j_1=1,2,\cdots a_1\}$. Given $\Omega(j_1)$, the total number of possible $\Omega(j_2)$ is a_2, and the set of all possible $\Omega(j_2)$ is represented by $\{\Omega(j_2), j_2=1,2,\cdots a_2\}$. Note that given

$\Omega(j_1)$ and $\Omega(j_2)$, the number of possible $\Omega(j_3)$ is 1 since the group G_3 is composed of all the left SSs that belong to neither G_1 nor G_2. In other words, j_3 is always 1. Thus, the service probability of the tagged SS at the channel state n is given as

$$\sigma_{S_n} = \begin{cases} \sum\limits_{k_1=0}^{h-1} \xi(\frac{h-k_1}{k_2}) \sum\limits_{j_1=1}^{a_1} \prod\limits_{i_1\in\Omega(j_1)} Pr(S_{i_1} > n) \sum\limits_{k_2=1}^{M-k_1} \sum\limits_{j_2=1}^{a_2} \prod\limits_{i_2\in\Omega(j_2)} Pr(S_{i_2} = n) \prod\limits_{i_3\in\Omega(j_3)} Pr(S_{i_3} < n) & n = 1, \cdots, 7 \\ 0 & n = 0 \end{cases} \tag{4.4}$$

Note that $\sigma_{S_0} = 0$ since the tagged queue is not allowed to transmit when the channel state of the tagged SS is at state 0, considering a high error bit rate at such a poor channel condition.

- **Inter-service Time**

The inter-service time for the tagged SS is defined as the average number of frames between two adjacent transmission opportunities for the tagged SS. An integrated Markov model is constructed to describe the tagged SS, where each state in the Markov model represents the current channel state of the tagged SS and whether the tagged queue obtains the chance of transmission in the current DL sub-frame. It consists of $2N + 1$ ($N = 7$ in the study) states as shown in Figure 4.2, where (n, s) and (n, w) represent that the tagged SS obtains and loses the chance of transmission, respectively, while its channel is at the state n. The transmission probability matrix is given by

$$\underline{\underline{Q}} = \begin{bmatrix} p_{00} & p_{01} & p_{01} & \cdots & p_{0N} & p_{0N} & \\ p_{10} & p_{11} & p_{11} & \cdots & p_{1N} & p_{1N} & \\ p_{10} & p_{11} & p_{11} & \cdots & p_{1N} & p_{1N} & \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ p_{N0} & p_{N1} & p_{N1} & \cdots & p_{NN} & p_{NN} & \\ p_{N0} & p_{N1} & p_{N1} & \cdots & p_{NN} & p_{NN} & \end{bmatrix} \cdot \begin{bmatrix} 1-\sigma_{S_0} & 0 & 0 & \cdots & 0 & 0 \\ 0 & \sigma_{S_1} & 0 & \cdots & 0 & 0 \\ 0 & 0 & 1-\sigma_{S_1} & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 1-\sigma_{S_N} & 0 \\ 0 & 0 & 0 & \cdots & 0 & 1-\sigma_{S_N} \end{bmatrix} \tag{4.5}$$

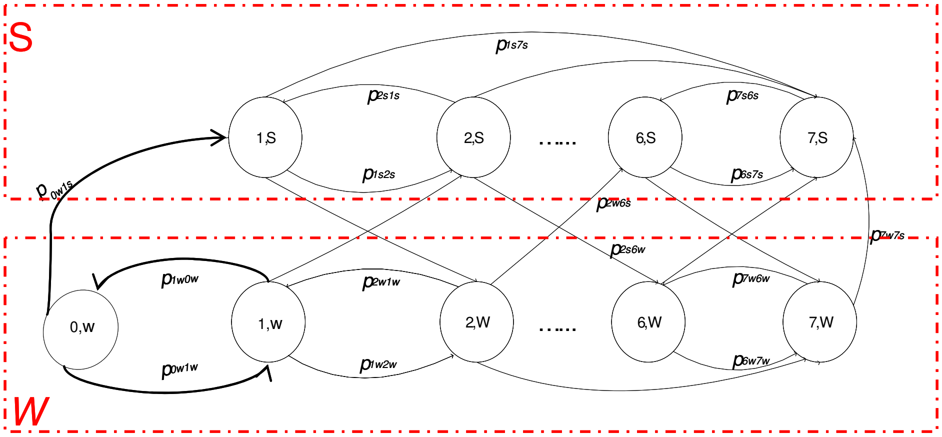


Figure 4.2: The Markov model for the tagged SS.

In order to derive the inter-service time, we group the states in Figure 4.2 into two states denoted as S and W, respectively. The states S and W represent the states where the tagged queue obtains and loses the chance of transmission, respectively.

The transition probabilities of the grouped states are given by

$$p_{sw} = \frac{\sum\limits_{n=1}^{7} [\theta(n,s) \sum\limits_{j=0}^{7} p_{ns,jw}]}{\sum\limits_{n=1}^{7} \theta(n,s)} \ , \qquad p_{ss} = 1 - p_{sw} \ , \tag{4.6}$$

$$p_{ws} = \frac{\sum\limits_{n=0}^{7} [\theta(n,w) \sum\limits_{j=1}^{7} p_{nw,js}]}{\sum\limits_{n=0}^{7} \theta(n,w)} \ , \qquad p_{ww} = 1 - p_{ws} \ , \tag{4.7}$$

where $\theta(n,s)$ is the steady-state probability of the state (n,s), and $p_{ns,jw}$ is the one-step transition probability from the state (n,s) to the state (j,w) $(n = 1, 2, \cdots, 7; j = 0, 1, \cdots, 7)$.

Let m denote the inter-service time (in the unit of frames), which is given by

$$E[m] = \sum_{i=1}^{\infty} i p_{sw} (p_{ww})^{i-1} p_{ws} = \frac{p_{sw}}{p_{ws}} \tag{4.8}$$

- **Achieved Goodput**

Goodput achieved by the tagged queue is defined as the average data rate (in the unit of bit per second) successfully lunched by the tagged queue. Let μ denote the number of PDUs successfully launched by the tagged queue during a transmission opportunity. The probability density function of μ is given as

$$Pr(\mu = i) = \begin{cases} (1-p)^i p & i = 0, 1, \cdots, L-1 \\ (1-p)^L & i = L \end{cases} \tag{4.9}$$

where p is the error probability of transmitting a PDU, L is the number of PDUs transmitted by the tagged queue during a DL sub-frame. The mean of μ is given by

$$E[\mu] = \sum_{i=0}^{L-1} i(1-p)^i p + L(1-p)^L \tag{4.10}$$

Thus, the goodput achieved by the tagged SS with cumulative ARQ is given by

$$G^C = \frac{E[u] \cdot B}{T \cdot (E[m] + 1)} bps \tag{4.11}$$

where $E[m]$ and $E[\mu]$ are the mean of the inter-service time and the number of PDUs successfully launched by the tagged queue in each transmission opportunity, respectively, T is the time duration of a frame, and B is the size of a PDU in the unit of bits.

- **Delivery delay of a PDU**

We evaluate the delivery of a PDU in the unit of frames, which is defined as the total number of frames lasting from the first transmission of a PDU to the frame during which this PDU is successfully received. Let N_P be the number of transmission attempts experienced by the tagged queue to successfully transmit a PDU, and m_i be the i-th inter-service time. Then, the delivery delay of a PDU can be given by

$$D_P^C = \sum_{i=1}^{N_P - 1} (m_i + 1) \tag{4.12}$$

Hence, the average delivery delay is given by

$$E[D_P^C] = E[\sum_{i=1}^{N_P-1} (m_i + 1)] = (E[N_P] - 1)(E[m] + 1) \tag{4.13}$$

The Calculation of $E[N_P]$ *:* We refer to L PDUs that are launched at each DL sub-frame as a transmission burst. Based on the principle of the cumulative ARQ discussed in Section 3.4, the delivery delay of a PDU is relative to its position at the transmission burst where its first transmission occurs. Note that whether or not a PDU is successfully transmitted depends not only on the successful transmission of itself but also on the successful transmission of all the previous PDUs in the transmission burst where its first transmission occurs. For a PDU whose first transmission occurs at the k-th position of a transmission burst, its transmission can be modeled by an absorbing Markov chain as shown in Figure 4.3, where the state i represents that this PDU is at the i-th position in a transmission burst, and the state 0 is the absorbing state representing that a PDU is transmitted successfully.

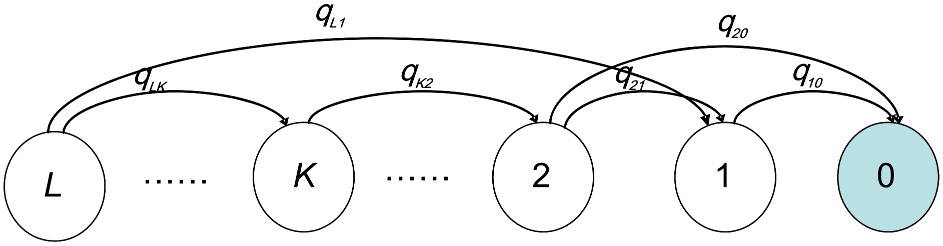


Figure 4.3: The transition diagram of a PDU.

The one-step transition probability matrix is give by

$$Q = \begin{bmatrix} q_{L,L} & q_{L,L-1} & \cdots & q_{L,0} \\ q_{L-1,L} & q_{L-1,L-1} & \cdots & q_{L-1,0} \\ \vdots & \vdots & \vdots & \vdots \\ q_{1,L} & q_{1,L-1} & \cdots & q_{1,0} \\ q_{0,L} & q_{0,L-1} & \cdots & q_{0,0} \end{bmatrix} \tag{4.14}$$

where q_{ij} is the transition probability from the state i to the state j, which is given by

$$q_{ij} = \begin{cases} 0 & i < j \\ (1-p)^i & j = 0 \\ (1-p)^{(i-j)}p & Otherwise \end{cases} \tag{4.15}$$

Hence, the expected number of transmission opportunities for successfully transmitting a PDU is equivalent to the average number of steps experienced by the PDU until to be absorbed, which is given by

$$E[N_P] = \Pi_0(I-R)^{-1}e \tag{4.16}$$

where Π_0 is the initial state vector, I is a $L \times L$ identity matrix, R is the matrix derived from the one-step transition probability Q by deleting the row and column corresponding to the absorbing state 0, and e is a column vector with all elements equal to 1.

Let ζ denote the probability that a PDU is located at the ith position in a transmission burst where its first transmission occurs. The initial state vector, $\Pi_0 = [\pi_1, \pi_2, \cdots, \pi_i, \cdots, \pi_L]$, is derived as follows:

$$\pi_i = Pr(\zeta = i) \quad = \sum_{j=1}^{L} [Pr(\zeta = i|\mu = j) \cdot Pr(\mu = j)] \tag{4.17}$$

$$Pr(\zeta = i|\mu = j) = \begin{cases} 1/j & i = L-j+1, \cdots, L \\ 0 & Otherwise \end{cases} \tag{4.18}$$

where μ denotes the number of PDUs successfully transmitted during each time the tagged queue obtains the chance of transmission, and its probability density function has been given by (4.9).

- **Delivery delay of an SDU/Packet**

We calculate the deliver delay of an SDU in the unit of frames, which is defined as the total number of frames lasting from the first transmission of the first PDU belonging to this SDU to the frame in which the last PDU of the SDU is successfully received. Let N_S be the number of transmission opportunities for successfully transmitting an SDU, and m_i be the ith inter-service time. Then, the delivery delay of an SDU is expressed by

$$D_S^C = \sum_{i=1}^{N_S-1} (m_i + 1) \tag{4.19}$$

Hence, the average delivery delay of an SDU is given by

$$E[D_S^C] = E[\sum_{i=1}^{N_S-1} (m_i + 1)] = (E[N_S] - 1)(E[m] + 1) \tag{4.20}$$

where $E[m]$ is the average inter-service time, and $E[N_S]$ is the expected number of transmission chances experienced by the tagged queue to successfully transmit an SDU, which is derived as follows.

The Calculation of $E[N_S]$: A two-queue model is developed to evaluate the delivery delay of each SDU at the tagged queue, where two logic queues, called the *transmission queue* (*tQ*) and the *waiting queue* (*wQ*), are devised. The *tQ* buffers the PDUs to be transmitted at the next transmission opportunity, while *wQ* buffers other PDUs. We assume that the failed PDUs have a higher priority during the next transmission opportunity. In other word, the PDUs in *tQ* are composed of all PDUs failed during the previous transmission and the PDUs from the *wQ* with the number of leftover quota out of L.

We observe the delivery of an arbitrary SDU in the tagged queue, which is referred to as the tagged SDU, and all PDUs belonging to the tagged SDU are referred to as the tagged PDUs. Let t_1 be the time instant at which the tagged queue wins the transmission opportunity and the first tagged PDU is transmitted in this opportunity. Let the time sequence $\{t_n : n > 1\}$ denote the following successive instants at which the tagged queue obtains the chances of transmissions, and $Loc_n \in \{0, 1, \cdots, L\}$ and $S_n \in \{0, 1, , \cdot, F\}$ represent the location of the first tagged PDU in the *tQ* and the total number of the successfully transmitted tagged PDUs observed at the instants

$\{t_n : n \geq 1\}$, respectively. The process $\{Loc_n, S_n : n = 1, 2, \cdots\}$ forms an absorbing embedded Markov chain on the state space $\{(0, 1, 2, \cdots, L) \times (0, 1, 2, \cdots, F)\}$, as shown in Figure 4.4, which represents the state transition of the tagged PDUs. The state $(0, F)$ is the absorbing state, representing that all the tagged PDUs are transmitted successfully. When the system reaches the absorbing state $(0, F)$, it means that the tagged SDU is transmitted successfully. The one-step transition probability matrix of this Markov chain is given by

$$\underline{\underline{W}} = [P_{ij,i'j'}] \quad i, i' \in \{0, 1, \cdots L\};\ j, j' \in \{0, 1, \cdots F\} \tag{4.21}$$

$$P_{ij,i'j'} = \begin{cases} (1-p)^{j'-j}p & i = 1, i' = 1, (j'-j) < L \\ (1-p)^{j'-j} & i = 1, i' = 1, (j'-j) = L \\ (1-p)^{j'-j} & i = 1, i' = 0 \\ (1-p)^{i-i'+j'}p & i \neq 1, i' \neq 0, (i-i'+j') < L \\ (1-p)^{i-i'+j'} & i \neq 1, i' \neq 0, (i-i'+j') = L \\ (1-p)^{i-1+j'} & i \neq 1, i' = 0 \end{cases} \tag{4.22}$$

where the element $P_{ij,i'j'}$ denotes the transition probability from the state (i, j) to the state (i', j').

Hence, the expected number of transmission opportunities for successfully transmitting an SDU is equivalent to the average number of steps experienced by the SDU until it is being absorbed, which is given by

$$E[N_S] = \Pi_0(I - R)^{-1}e \tag{4.23}$$

where Π_0 is the initial state vector, I is the $(L+1)(F+1) \times (L+1)(F+1)$ identity matrix, R is the matrix derived from the one-step transition probability $\underline{\underline{W}}$ by deleting the row and column corresponding to the absorbing state $(0, F)$, and e is a column vector with all elements equal to 1.

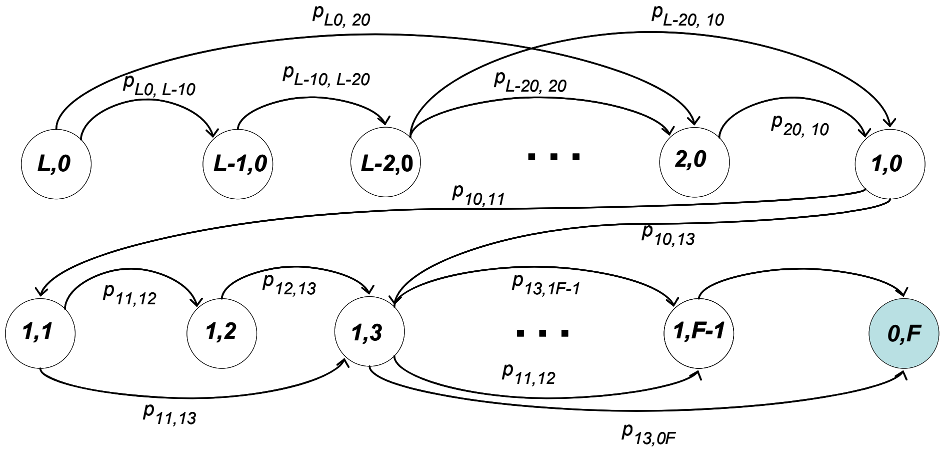


Figure 4.4: The state transition diagram for the delivery of an SDU.

In order to calculate $E[N_S]$, we need to know the initial state vector Π_0, which is analyzed as follows. Firstly, the PDUs at the wQ are indexed with the mod of L. Let the time sequence $\{\delta_n : n > 0\}$ denote the successive time instants that the tagged queue obtains the transmission opportunities. Let random variable $\{\phi_n : n = 1, 2, \cdots\}$ be the index of the head-of-line (HoL) PDU at the wQ observed at $\{\delta_n : n > 0\}$. The process $\{\phi_n : n = 1, 2, \cdots\}$ forms an embedded Markov chain on the state space $\{1, 2, \cdots, L\}$, as shown in Figure 4.5.

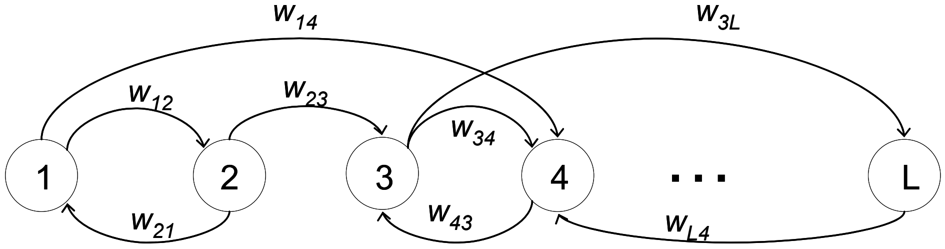


Figure 4.5: The state transition diagram of HOL PDUs at the waiting queue.

The state transition probability of this Markov chain depends on the number of PDUs at the tQ being successfully transmitted at each transmission opportunity, which equivalently depends on the number of PDUs failed at the previous transmission. Therefore, the one-step transition probability matrix is given by

$$\Omega = \begin{bmatrix} w_{11} & w_{12} & \cdots & w_{1L} \\ w_{21} & w_{22} & \cdots & w_{2L} \\ \vdots & \vdots & \vdots & \vdots \\ w_{L1} & w_{L2} & \cdots & w_{LL} \end{bmatrix} \tag{4.24}$$

where w_{ij} is the transition probability from the state i to the state j, which is given by

$$w_{ij} = \begin{cases} (1-p)^{j-i}p & j > i \\ (1-p)^{L-i+j}p & j < i \\ (1-p)^{L} & i = j \end{cases} \tag{4.25}$$

Based on the one-step transition probability matrix, the steady-state probability $h_i = \lim_{n \to \infty} pr[\phi_n = i]$ $(i = 1, 2, \cdots, L)$ is derived from the balance equations:

$$\begin{cases} H = H\Omega \\ \sum_{1}^{L} h_i = 1 \end{cases} \tag{4.26}$$

where H denotes the steady-state probability vector, and Ω is the one-step transition probability matrix.

The state transition from the state i to the state j determines the occurrence of some specific initial states. For instance, the transition from the state 1 to the state L, which is due to the successful transmission of $(L-1)$ PDUs at the previous transmission. In other words, the initial state $(2, 0)$, $(F+2, 0)$, $\cdots$, $((\lfloor (L-1)/F \rfloor \cdot F + 2), 0)$ occurs one time simultaneously. Since the transition from the state i to the state j occurs with the probability $h_i w_{ij}$, the probability by which the corresponding initial states occur can be obtained accordingly. Therefore, the initial state vector Π_0 can be derived based on the steady-state probability vector H and the one-step transition probability matrix Ω. When the obtained initial state probability vector Π_0 is obtained, the average number of steps for an SDU to absorption is derived from (4.23), and SDU delivery delay is derived from (4.20).

4.2.2 Performance Analysis with Selective ARQ

Service probability and inter-service time depends on the resource allocation and scheduling framework. They are independent of the ARQ mechanism adopted in the system. Therefore, with the selective ARQ, the service probability and inter-service time are derived by (4.4) and (4.8), respectively, same as that with the cumulative ARQ.

- **Resource Utilization**

Let τ be the resource utilization achieved by the system, which is defined as the number of information bits carried by an OFDM symbol. When a high modulation level is used, the number of information bits carried by an OFDM symbol is large, which yields a high resource utilization, and vise versa. Therefore, resource utilization is an important metric to evaluate the performance of the proposed framework, which can be obtained as

$$E[\tau] = \sum_{i=1}^{M} \frac{\sum_{n=1}^{7} \theta_i(n,s)}{\sum_{j=1}^{M} \sum_{n=1}^{7} \theta_j(n,s)} \sum_{k=1}^{7} I_k \frac{a_k \theta_i(k,s)}{\sum_{n=1}^{7} a_n \theta_i(n,s)} \tag{4.27}$$

where M is the total number of SSs in the network, $\theta_i(n,s)$ is the steady-state probability of the state (n,s) for SS_i, I_k is the information bits carried by an OFDM symbol when the channel state is at k, which is given in Table 3.1, and a_n is the required OFDM symbols to transmit L PDUs at the tagged queue when the channel state is n.

- **Achieved Goodput**

Let η denote the number of PDUs successfully launched by the tagged queue during each transmission opportunity. The mean of η is given by

$$E[\eta] = \sum_{i=0}^{L} \left[i \binom{L}{i} (1-p)^i p^{L-i} \right] = L \cdot (1-p) \tag{4.28}$$

where L is the number of PDUs transmitted by the tagged queue each time it obtains the transmission opportunity.

Goodput achieved at the tagged queue is defined as the average data rate (in unit of bit per second) successfully launched by the tagged queue, and is given by

$$G^S = \frac{E[\eta] \cdot B}{T \cdot (E[m]+1)} = \frac{L \cdot (1-p) \cdot B}{T \cdot (E[m]+1)} \; bps \tag{4.29}$$

where p is the error probability of transmitting a PDU, B is the size of a PDU in the unit of bits, T is the time duration of a MAC frame, and $E[m]$ is the mean of the inter-service time, which is derived from (4.8).

- **Delivery Delay of a PDU**

The average number of transmission/retransmission for a PDU is given by

$$\sum_{n=1}^{\infty} np^{n-1}(1-p) = 1 + \frac{p}{1-p} \tag{4.30}$$

where p is the error probability of transmitting a PDU.

Thus, the average delivery delay of a PDU (in unit of frame) is given by

$$E[D_P^S] = \frac{p}{1-p}(E[m]+1) \tag{4.31}$$

where $E[m]$ is the mean of the inter-service time.

- **Delivery Delay of an SDU/Packet**

The deliver delay of an SDU is defined as the number of MAC frames counted from the launching of the first PDU of the SDU until the successful recipient of the last PDU of the SDU. Let the random variable N_S denote the number of transmission opportunities experienced by a queue to successfully transmit an SDU, and m_i be the ith inter-service time. The mean of the delivery delay of an SDU is given by

$$E[D_S^S] = E[\sum_{i=1}^{N_S-1} m_i + 1] = (E[N_S] - 1)(E[m] + 1) \qquad (4.32)$$

where $E[N_S]$ is the expected number of transmission opportunities experienced by the tagged queue to successfully transmit an SDU, and $E[m]$ is the average inter-service time, which has been derived from (4.8). Thus, we need to obtain $E[N_S]$ in order to derive $E[D_S^S]$.

The Calculation of $E[N_S]$: Similar to the calculation of $E[N_P]$, we develop a two-queue model to evaluate the delivery of SDUs at the tagged queue of the BS, where two logic queues, called the *transmission queue* (tQ) and *waiting queue* (wQ), are devised, as shown in Figure 4.6. The tQ buffers the PDUs to be transmitted at the next transmission opportunity, while wQ buffers other PDUs. We assume that the failed PDUs have a higher priority during the next transmission opportunity. In other word, the PDUs in tQ are composed of all PDUs failed during the previous transmission and the PDUs from the wQ with the number of leftover quota out of L.

Let an arbitrary SDU in the *tagged* queue be referred to as the *tagged* SDU, all PDUs belonging to the *tagged* SDU be referred to as the *tagged* PDUs, t_1 be the time instant at which the tagged queue wins a transmission opportunity and launches the first tagged PDU, the subsequent instants at which the tagged queue obtains the transmission opportunity be denoted as $\{t_n : n > 0\}$, and the random variables $A_n \in \{0, 1, \cdots F\}$ and $f_n \in \{0, 1, \cdots F\}$represent the number of the tagged PDUs in the tQ and the wQ observed at the instants $\{t_n : n > 0\}$, respectively. The process

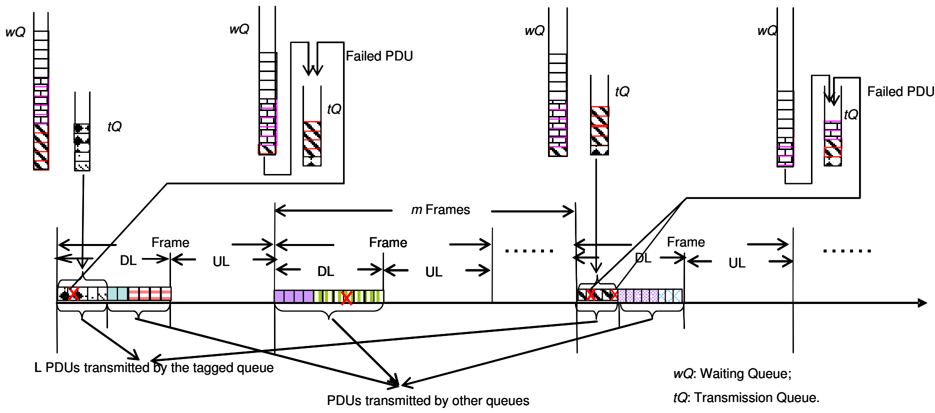


Figure 4.6: The two-queue model for the delivery of SDUs.

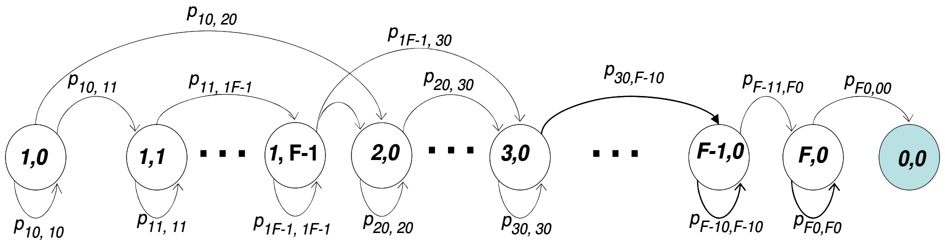


Figure 4.7: The state transition diagram of the absorbing Markov chain.

$\{A_n, f_n : n = 1, 2, \cdots\}$ forms an absorbing embedded Markov chain on the state space $\{(0, 1, 2, \cdots F) \times (0, 1, 2, \cdots F)\}$, as shown in Figure 4.7, which represents the state transition of the tagged PDUs. The state $(0, 0)$ is the absorbing state, representing that all the tagged PDUs are launched successfully. When the system reaches the absorbing state $(0, 0)$, the tagged SDU is completely transmitted. The one-step transition probability matrix of this Markov chain is given by

$$P = [p_{ij,\ i'j'}] \quad i, j, i', j' \in \{0, 1, 2, \cdots, F\} \tag{4.33}$$

where P is a $(F + 1)(F + 1)\times(F + 1)(F + 1)$ matrix, and the element $p_{ij,i'j'}$ denotes the transition probability from the state (i, j) to the state (i', j').

Let x_{in} represent the number of the tagged PDUs moving to the tQ queue from the wQ queue due to the successful transmission of some PDUs in the current transmission opportunity, and x_{out} denote the number of the tagged PDUs successfully launched in the current transmission opportunity. Therefore, the transition from the state (i, j) to the state (i', j') specifically determines a pair of (x_{in}, x_{out}), which is given by

$$x_{in} = j - j' \ , \qquad x_{out} = (i + j) - (i' + j') \tag{4.34}$$

Hence, the transition probability $p_{ij,i'j'}$ is given by

$$p_{ij,\ i'j'} = \begin{cases} [\binom{i}{x_{out}}(1-p)^{x_{out}} p^{i-x_{out}}][\binom{L-i}{x_{in}-x_{out}}(1-p)^{x_{in}-x_{out}} p^{(L-i)-(x_{in}-x_{out})}] & (i,j,i',j') \in T_1 \\ [\binom{i}{x_{out}}(1-p)^{x_{out}} p^{i-x_{out}}][\sum\limits_{m=x_{in}-x_{out}}^{L-i} \binom{L-i}{m}(1-p)^{m} p^{(L-i)-m}] & (i,j,i',j') \in T_2 \\ \binom{i}{x_{out}}(1-p)^{x_{out}} p^{i-x_{out}} & (i,j,i',j') \in T_3 \\ 0 & (i,j,i',j') \in T_4 \end{cases} , \tag{4.35}$$

$$T_1 := \{(i,j,i',j') \in Q \mid j, j' \neq 0, x_{in} \geq x_{out}\} \ , \qquad T_2 := \{(i,j,i',j') \in Q \mid j \neq 0, j' = 0, x_{in} \geq x_{out}\} \ ,$$
$$T_3 := \{(i,j,i',j') \in Q \mid x_{in} < x_{out}\} \ , \qquad T_4 := \overline{Q} \ ,$$
$$Q := \{(i,j,i',j') \in T \mid 0 \leq x_{in} < j, 0 \leq x_{out} \leq min(x_{in}, i);\ x_{in} = j, 0 \leq x_{out} \leq i\} \ ,$$
$$T := \{(i,j,i',j') \mid (j,j) \in S, (i',j') \in S\} \ , \qquad S := \{(i,j) \mid 0 \leq i \leq F, 0 \leq j \leq F - i\}$$

where T_1, T_2,and T_3 are the subsets of Q, T_4 is the complementary set of Q in T, Q is a subset of S, and S and T are the state space and transition space of the absorbing Markov chain, respectively. Hence, the expected number of transmission opportunities required to successfully transmit the tagged SDU is equivalent to the average number of steps to be absorbed for the tagged SDU, which is given by

$$E[N_s] = \Pi_0 (I - R)^{-1} e \tag{4.36}$$

where Π_0 is the initial state vector, I is the identity matrix, R is the matrix derived from the one-step transition probability by deleting the row and column corresponding to the absorbing state $(0, 0)$, and e is a column vector with all elements equal to 1. In order to calculate $E[N_S]$, we need to know the initial state vector Π_0, which is analyzed as follows.

First, the PDUs at the wQ queue is indexed with the mod L. Let the successive time instant the tagged queue obtains the transmission opportunities be denoted as $\{t_n : n > 0\}$ and random variable $\phi_n \in \{1, 2, \cdots L\}$ be the index of the HoL PDU at the wQ queue observed at $\{t_n : n > 0\}$. The process $\{\phi_n : \ n > 0\}$ forms a embedded Markov chain on the state space $\{1, 2, \cdots, L\}$, as shown in Figure 4.8.

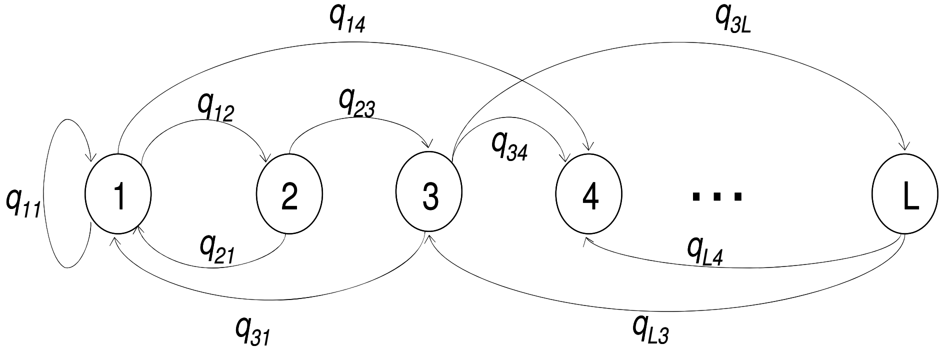


Figure 4.8: The state transition diagram of the HOL PDU at Waiting Queue.

The state transition probability of this Markov chain is determined by the number of PDUs at the tQ queue being launched at each transmission opportunity, which is equivalently the number of PDUs failed in each transmission opportunity. Therefore, the one-step transition probability from the state i to the state j, q_{ij}, is given by

$$q_{ij} = \begin{cases} \binom{L}{j-i}(1-p)^{j-i}p^{L-(j-i)} & j > i \\ (1-p)^L & j = i \\ \binom{L}{L+j-i}(1-p)^{L+j-i}p^{i-j} & j < i \end{cases} \tag{4.37}$$

Based on the one-step transition probability, the steady-state probability $h_i = \lim_{n\to\infty} Pr(\phi_n = i)(i = 1, 2, \cdots, L)$ can be derived from the balance equations. The state transition from the state i to the state j determines the occurrence of some specific initial states. For instance, the transition from the state 1 to the state L, which is due to successful transmission of $(L-1)$ PDUs at the previous transmission opportunity, implies that the initial state $(F,0)$ occurs $\lfloor (L-1)/F \rfloor$ times, and the initial state $(F-1,1)$ occurs once concurrently. Since the transition from the state i to the state j occurs with the probability $h_i q_{ij}$, the probability that the corresponding initial states occur can be obtained accordingly. Therefore, the initial state vector Π_0 can be derived based on the derived steady-state probability h_i and the one-step transition probability matrix q_{ij}. When the initial state probability vector Π_0 is obtained, the average number of steps for an SDU to be absorbed and the average delivery delay of SDU can be derived from (4.36) and (4.32), respectively.

4.3 Simulation Results

Extensive simulations are conducted using MATLAB to demonstrate the performance of the proposed framework in terms of SDU delivery delay, achieved goodput, and resource utilization. The impact of two parameters, h and L, on the performance metrics are illustrated by extensive simulations (denoted as Sim), which also verify the accuracy of the analytical model (denoted as Ana). We repeat the simulation 50 times with different random seeds and calculate the average value.

In the simulation, the Rayleigh fading channel model is adopted, and the total number of SSs is 5 with the average signal-to-noise ratio (ASNR) of 5, 10, 15, 20 and 25dB, respectively. The other simulation parameters are given in Table 4.1.

Figures 4.9 – 4.13 illustrate the performance of the proposed framework with the cumulative ARQ. Figure 4.9 shows the SDU delivery delay versus L for different p values, respectively. It can be seen that the SDU delivery delay decreases with the increase of L. With a larger L, more PDUs are transmitted when the tagged queue obtains a transmission opportunity, which leads to more resources allocated to the tagged queue. As a result, a smaller average SDU delivery delay is achieved. It is also seen that with a smaller p, the SDU delivery delay decreases due to the increase of the

Table 4.1: Simulation Parameters

MAC frame duration	2.5 ms
DL/UL sub-frame duration	1.25 ms/1.25 ms
OFDM symbol duration	23.8μs
Doppler frequency	15 Hz
Bandwidth	10 MHz
The number of SSs	5
Indices of SSs	1, 2, 3, 4, and 5
The number of SSs selected in each DL sub-frame	2
Tagged SS	SS_3
The average SNR of each SS	5, 10, 15, 20, and 25 dB
Size of a PDU	1536 $bits$
Size of an SDU	15 $PDUs$
Assigned resource during each chance of transmission	5 $PDUs$
Error probability of transmitting a PDU	0.01

probability of successfully transmitting a PDU.

Figure 4.10 shows the goodput of the tagged queue versus L for different PDU error probability p. With a larger L, more resources are allocated to the tagged queue. Thus, the achieved goodput is larger. From Figures 4.9 - 4.10, it can be seen that the simulation and analysis results match very well, which verifies the accuracy of the analytical model.

Figures 4.11 – 4.13 illustrate the impacts of h on the performance metrics in terms of SDU delivery delay, inter-service time, and goodput. From Figures 4.11 and 4.12, it is observed that both SDU delivery delay and inter-service time decrease with the increase of h. With a larger h, SSs are visited more frequently and have more chances

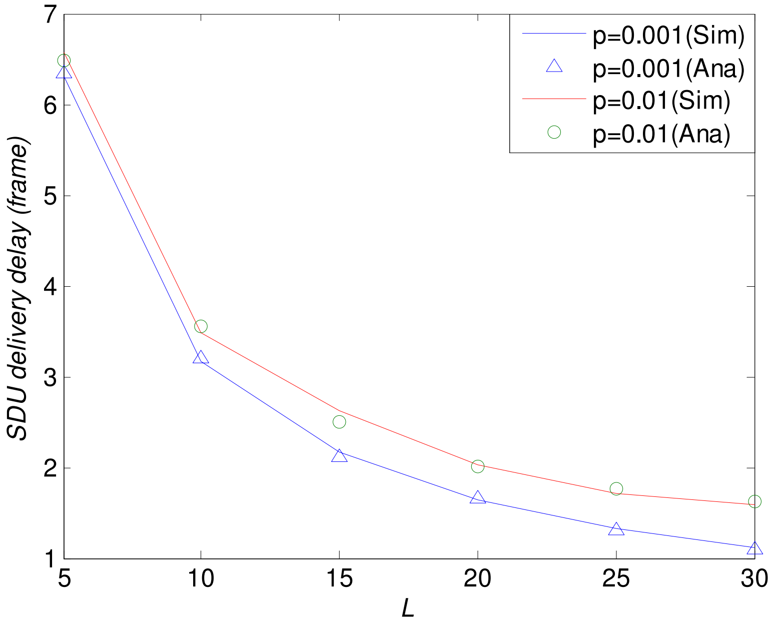


Figure 4.9: SDU delivery delay versus L with different p.

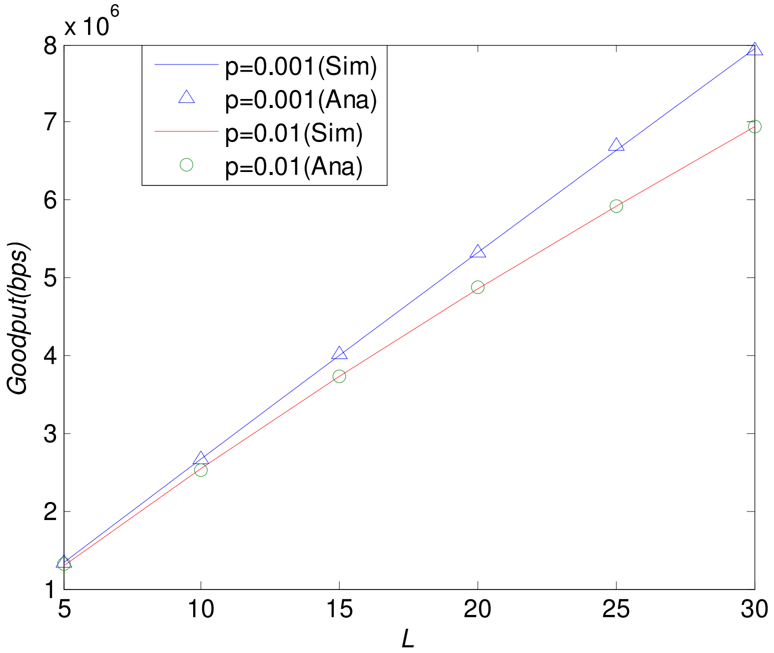


Figure 4.10: Goodput versus L for different p.

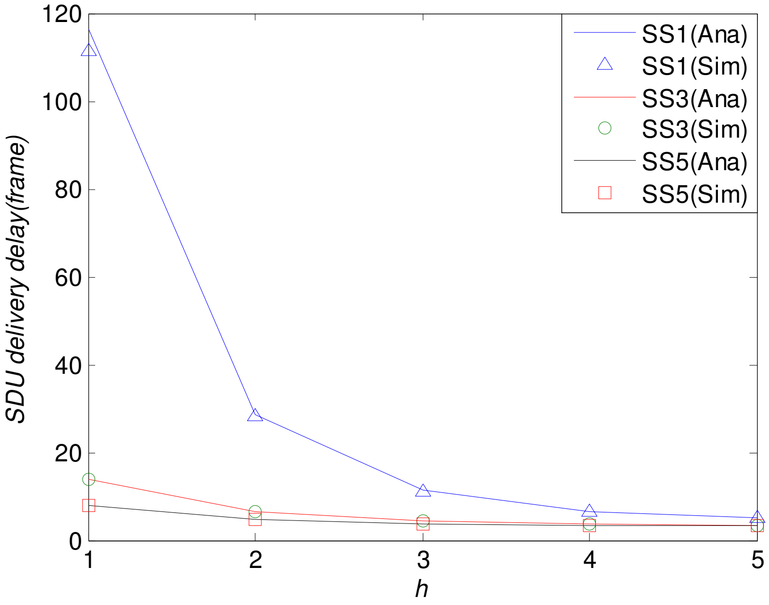


Figure 4.11: SDU delivery delay versus h.

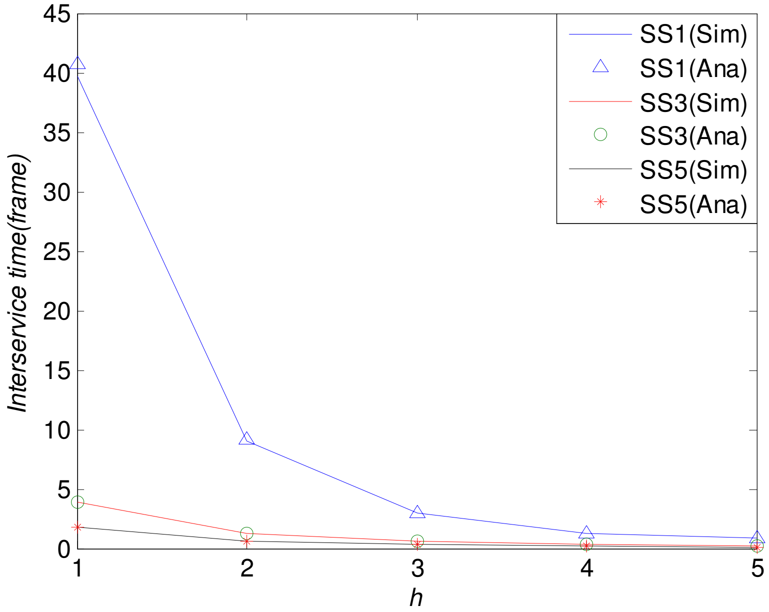


Figure 4.12: Inter-service time versus h.

of transmissions, which leads to a shorter delivery delay and inter-service time. The opportunistic scheduling is the specific case with h=1. In this case, SS1, with the worst

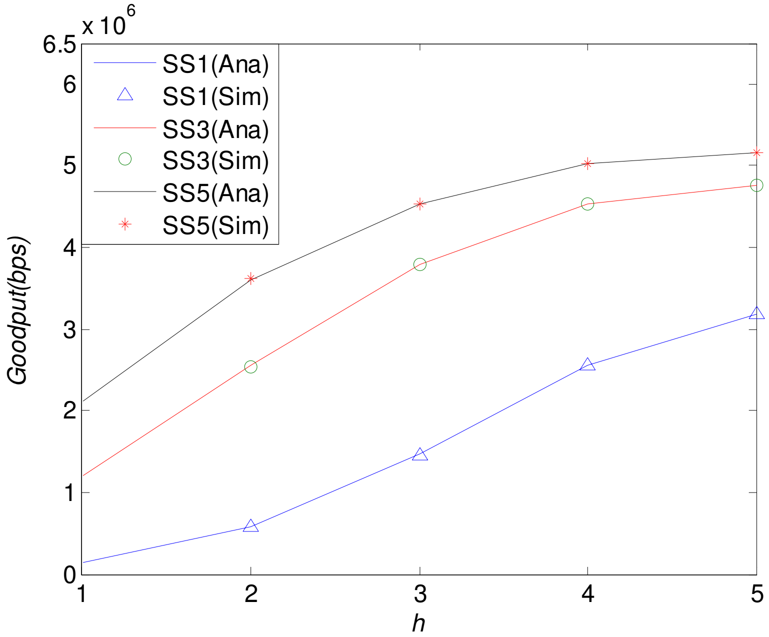


Figure 4.13: Goodput versus h.

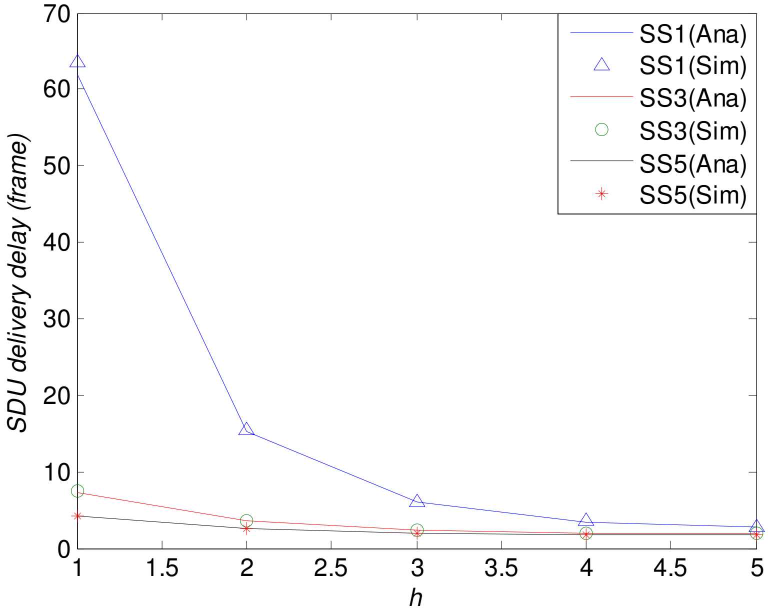


Figure 4.14: SDU delivery delay versus h.

channel condition, is starved for a long time. Its delivery delay is much larger than that of other SSs. Figure 4.13 shows the obtained goodput of SSs for different h. With the

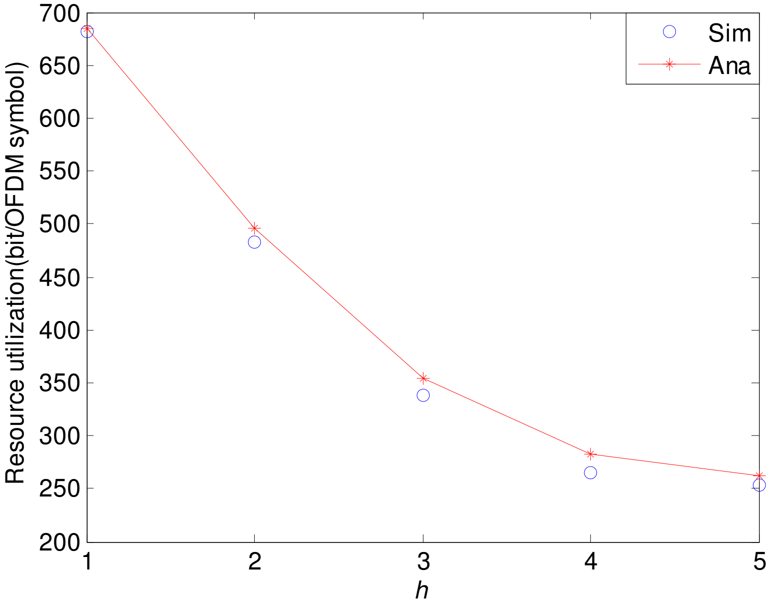


Figure 4.15: Resource Utilization versus h.

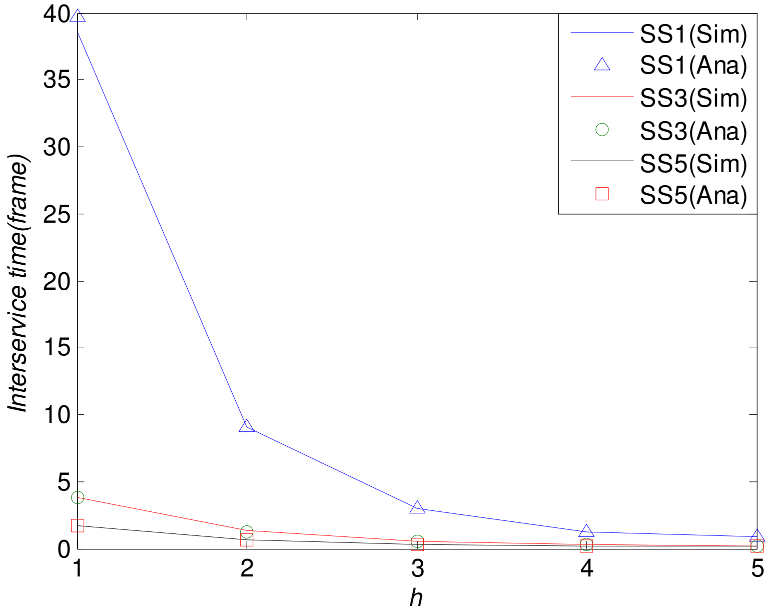


Figure 4.16: Inter-service time versus h.

increase of h, inter-service time of each SS decreases. Each SS obtains more chances of transmissions, leading to a higher throughput.

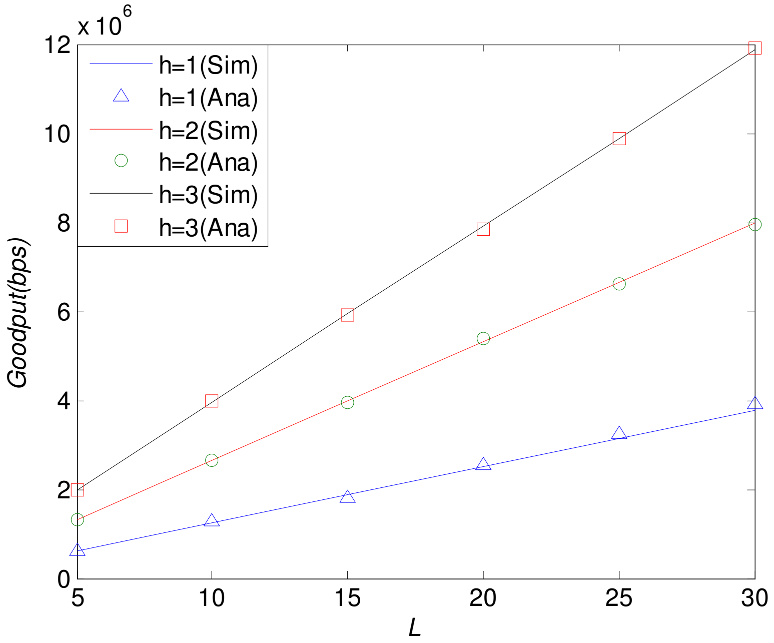


Figure 4.17: Goodput versus L.

Figures 4.14 – 4.17 illustrate the performance of the proposed framework with the selective ARQ. Figure 4.14 shows the impacts of the parameter h on the SDU delivery delay for different SSs. It can be seen that the SDU delivery delay decreases with the increase of h. When h is 1, one SS is selected to obtain the transmission opportunity at each DL sub-frame. Thus, SS1, which is subject to the worst channel condition, experiences a quite long delivery delay. With the increase of h, SS1 gets more chances of transmissions. Therefore, its SDU delivery delay decreases accordingly.

Figure 4.15 shows the impact of h on the resource utilization. It is observed that the resource utilization decreases with the increase of h. With a large h, some SSs with poor channel conditions can achieve more chances of transmissions, which leads to the use of a low modulation and coding level. As a result, a low resource utilization is obtained. From Figures 4.14 and 4.15, it can be seen that the parameter h plays a key role in manipulating the SDU delivery delay and the resource utilization. A small h leads to a better resource utilization, but it also leads to a large difference of SDU delivery delay between SSs with different channel conditions and a longer SDU delivery delay for the SS with a poor channel condition, and vice versa.

Figure 4.16 shows the relation between the inter-service time and h. It is observed that the inter-service time of each SS decreases with the increase of h. When h is 1, the inter-service time of SS1 is almost 22.3 times as that of SS5.

Figure 4.17 shows the relation between the achieved goodput and assigned bandwidth L with different h. It is observed that given a specific h, the goodput requirement can be achieved by manipulating a proper L. With the increase of L, the achieved goodputs of SSs increase accordingly. With a fixed L, the achieved goodput is impacted by the selection of h. With a larger h, SSs can obtain a higher goodput.

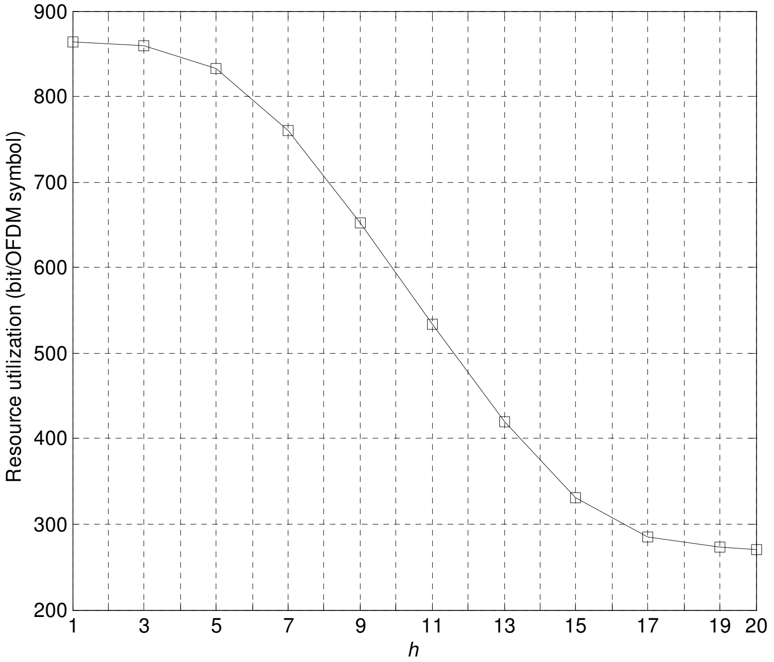


Figure 4.18: Resource utilization versus h.

The simulation results with a larger number of SSs are given in Figures 4.18 –4.19, where the total number of SSs is 20 which are divided into four groups with ASNRs of 10, 15, 20, and 25 dB, respectively. Each group includes 5 SSs with the same ASNR. Similar to Figures 4.14 – 4.15, it can be seen that h is a key parameter balancing the

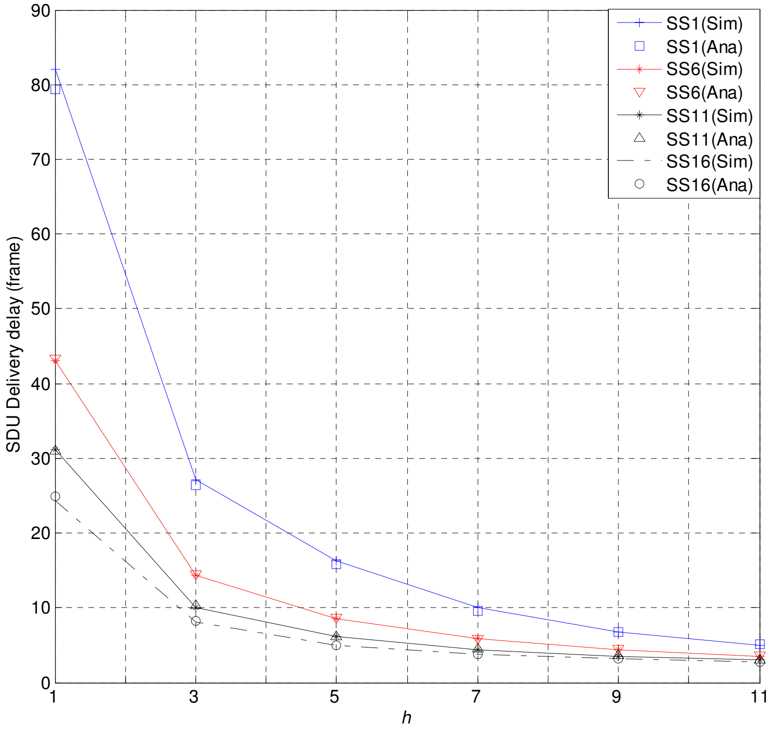


Figure 4.19: SDU delivery delay versus h.

delivery delay of SSs and resource utilization. With a small h, the network can achieve a high resource utilization, but the SSs with bad channel conditions experience a long delivery delay, and vise versa.

Figures 4.9 – 4.19 illustrate the impact of h and L on some important performance metrics. h and L are critical for the performance of the proposed framework. The setting of h is based on the requirement of the tradeoff between the SDU delivery delays and the resource utilization. The possible h is a value in the set of [1,2,...,M], where M is the total number of SSs in the network. Since the number of possible h equals M, the SDU delivery delay and resource utilizations corresponding to each possible h can be obtained using the analysis given in Section 4.2. When the sets of possible SDU delivery delays and resource utilizations are obtained, the parameter h can be selected based on the requirement of the SDU delivery delay and resource utilization. After h is set, L can be set by using analytical result given in (4.11) or (4.29), based on the throughput requirement of each SS. Furthermore, the parameters h and L decide the amount of resource assigned to nrtPS applications at each MAC frame, which can provide a useful guideline for the connection admission control such that the system will not be overloaded.

4.4 Discussion and Summary

A simple yet efficient resource allocation and packet scheduling framework has been proposed for nrtPS applications in IEEE 802.16 networks, where two ARQ mechanisms at the MAC layer and AMC technique at the PHY layer are jointly considered. An analytical model has been developed to provide meaningful guidelines to select appropriate parameters for satisfying the throughput requirement and initiating a graceful compromise between the delivery delay and the resource utilization for nrtPS applications. Extensive simulations have been conducted to demonstrate the effectiveness and efficiency of the proposed framework and verify the accuracy of the analytical model. Although the slow fading channel is considered in this study, the analytical model on the delivery delay, goodput, and resource utilization is also valid for other types of channels. The only difference is the derivation of the state transition matrix given in (3.4). Similar Markov models for other types of fading channels have been discussed in [86] [87] [88]. In addition, The proposed design framework is focused on the average throughput requirements and the tradeoff between the delivery delay and resource utilization. In terms of packet arrival, the only assumption in the analytical model is that when a queue is visited, it has packets waiting for transmission.

Chapter 5

Scheduling for Best Effort Service

5.1 The Weighted Proportional Fair Scheduling Scheme

Many scheduling schemes have been proposed to deal with different service types based on their intrinsic characteristics. For BE service, the main concern is to achieve a satisfying fairness and throughput performance. One of the most effective schemes is the proportional fairness scheduling, which can provide a good balance between the system throughput and fairness. However, the fairness in this scheduling scheme is defined on the basis that the traffic load and allocated resources are identical/homegeneous among all end users. This design requirement, however, is insufficient in IEEE 802.16 networks, where an SS could be a residential house, a mobile user, or an office building providing the Internet service to many customers. Multiple types of SSs is a unique feature of IEEE 802.16 networks, compared with the conventional wireless communication systems such as the 3G cellular systems with all the end users as individual cell phones and handsets. Due to the multiple types of SSs, each SS may submit a much different long-term traffic load and demand patterns. For instance, the BE traffic load for an SS of office building could be much higher than that for an SS of resident house during the day time, while residence houses could be the major capacity consumers during the late evening. The conventional scheduling schemes based on proportional fairness have focused on equivalently allocating the available resources among the users, which are rather efficient when the traffic demands of all users are homogeneous, yet are not suitable for IEEE 802.16 networks due to the potential heterogeneity among SSs in terms of traffic patterns and demands. Such a unique feature of IEEE 802.16 networks has posed new challenges on the design of an efficient resource allocation and scheduling scheme, and is particularly distinguished in the BE service which is subject to no any delay and minimum bandwidth requirement. It is clear that the ignorance of this fact would certainly lead to inefficiency of system operation, which could easily lead to a performance degradation due to the lack of consideration of each SS's potential traffic pattern. An efficient scheduling scheme for BE service should be able to allocate the available resources among different SSs in an adaptive and flexible way, such that network operators can freely perform the bandwidth allocation according to some historical behavior and statistic data of traffic pattern for each SS.

In this chapter, we propose the weighted proportional fair (WPF) scheduling scheme for BE service in IEEE 802.16 Networks for achieving flexible resource allocation according to channel conditions and traffic patterns. Furthermore, an analytical model is developed to study some important performance metrics, such as the spectral efficiency, throughput, resource utilization, and fairness, and evaluate the impacts of the weights and channel conditions on these performance metrics. The analytical model considers two scenarios: one is based on the Rayleigh fading channel; the other considers the AMC technique. Extensive simulations are conducted to illustrate the efficiency of the proposed scheme and verify the accuracy of the analytical model, where the impacts of the weights on performance metrics under the different channel conditions are further investigated.

Considering the different traffic demands among SSs, we introduce weight in the WPF to adjust the priority of different SSs to be selected for service. Meanwhile, instantaneous channel conditions should be considered to exploit the multi-user channel diversity. Therefore, the criterion by which the WPF scheduling scheme selects an SS at the beginning of each MAC frame and provides the resources available for BE service to this SS is given by

$$i^* = \arg\max_i X_i \tag{5.1}$$

$$X_i = w_i \frac{\gamma_i}{\overline{\gamma_i}} \tag{5.2}$$

where X_i represents the weighted relative channel condition of SS_i, w_i is the weight for SS_i, while $\overline{\gamma_i}$ and γ_i are the average and instantaneous channel condition for SS_i, respectively.

We define (5.2) as the preference metric of WPF scheduling scheme. By averaging out the long-term channel condition $\overline{\gamma_i}$ in the preference metric, WPF improves the short-term fairness. Meanwhile, WPF can achieve a high system throughput by exploiting multi-user channel diversity. The weight w_i reflects the BE traffic demand of SS_i. For easy implementation, we set the weight of each SS in such a way that the ratio of weights for SSs equals to the ratio of their traffic demands. That is, $\frac{w_i}{w_j} = \frac{D_i}{D_j}$ $(i, j = 1, 2, \cdots, M)$, where D_i is the BE traffic demand of SS_i. To implement the WPF scheduling scheme, BS should have the knowledge of the channel state information of each SS. In IEEE 802.16 networks, the uplink channel quality indication channel (UL CQICH) is allocated for SSs to feedback channel state information. In addition, some SSs in IEEE 802.16 networks are stationary, such as office buildings and residence houses. This feature can largely decrease the frequency of channel state feedback since the channel conditions of these SSs are less fluctuant. Thus, the overhead of implementation is reduced. Based on the channel condition of each SS and the parameters w_i $(i = 1, 2, \cdots, M)$, a preference metric vector $\underline{X} = [X_1, X_2, \cdots, X_M]$ is maintained at the scheduler at the BS. At the beginning of each frame, $\underline{X}$ is updated according to (5.2), while the BS selects an SS with the largest value of preference metric according to (5.1) and allocates the resource available for BE service at this frame to this SS.

By manipulating the weight w_i in the preference metric, the proposed scheme not only can provide flexible resource allocation among all SSs, but also can achieve a satisfying fairness performance due to the consideration of the traffic demands of different SSs. The following is to quantify the relation between the weights, channel conditions, and important performance measurements.

5.2 Performance Analysis

In this section, an analytical model is developed to investigate some important performance metrics, such as the service probability, spectral efficiency, throughput, and resource utilization. The analytical model includes two scenarios. One is based on the Rayleigh fading channel. The other is based on the N-state Markov channel model with the consideration of AMC technique. The notation used in the rest of the paper are listed in Table 5.1.

5.2.1 Performance Analysis Based on the Rayleigh Fading Channel

In thus subsection, we study some important performance metrics based on the Rayleigh flat fading channel [89–91], which provide us an upper bound on spectral efficiency and throughput.

- Service Probability for SS_i

The service probability for SS_i is defined as the probability that SS_i is selected for service at an arbitrary MAC frame when the system is stable. Based on (5.1), an SS with the largest value of preference metric is selected for service in each MAC frame. Therefore, the distribution of preference metric value of each SS plays a key role to analyze the service probability. Let function $l_i(\cdot)$ be the probability density

Table 5.1: Table of Notations

M	The total number of SSs in the network
$\overline{\gamma}_i$	The average SNR of SS_i
γ_i	The instantaneous SNR of SS_i
w_i	Weight of SS_i
D_i	BE traffic demand of SS_i
X_i	The value of Preference metric of SS_i
π_i	Service probability of SS_i
ψ	index of SS selected for service during a frame
ς_i	Spectral efficient for SS_i
ς	System spectral efficient
τ_i	Resource utilization for SS_i
τ	System resource utilization
Th_i^R	The throughput of SS_i with the Rayleigh fading channel
Th_i^A	The throughput of SS_i with the AMC
Th^R	The system throughput with the Rayleigh fading channel
Th^A	The system throughput with the AMC

function (p.d.f.) of X_i, where X_i is the preference metric value of SS_i given in (5.2). We define a function g_i as $X_i = g_i(\gamma) = w_i \frac{\gamma_i}{\overline{\gamma}_i}$. Based on the p.d.f. of dependent random variables [92], the p.d.f. of X_i is given by

$$\begin{aligned} l_i(X_i = x) &= \left| \frac{1}{g_i^{'}(g_i^{-1}(x))} \right| f_i(g_i^{-1}(x)) \\ &= \frac{1}{w_i} \, e^{-\frac{x}{w_i}} \end{aligned} \tag{5.3}$$

where $|\cdot|$ denotes the determinant of a matrix, function $f_i(\cdot)$ is the p.d.f. of γ_i, which is given in (3.2), $g_i^{-1}(\cdot)$ is the inverse function of $g_i(\gamma) = w_i \frac{\gamma}{\overline{\gamma}_i}$, and $g_i^{'}(\cdot)$ is the derivative of function $g_i(\cdot)$.

Let ψ be the index of the SS selected for service at a MAC frame. The service probability for SS_i is given by

$$\begin{aligned}\pi_i = Pr\{\psi = i\} &= \int_0^\infty l_i(x)(\prod_{j=1,j\neq i}^{M} \int_0^x l_j(y)dy)dx \\ &= \int_0^\infty \frac{1}{w_i} e^{\frac{-x}{w_i}} (\prod_{j=1,j\neq i}^{M} (\int_0^x \frac{1}{w_j} e^{\frac{-y}{w_j}} dy))dx\end{aligned} \tag{5.4}$$

where function $l_i(\cdot)$ is p.d.f of preference metric value for SS_i, which is given in (5.3).

Expanding the expression of π_i using

$$\prod_{j=1,j\neq i}^{M} (1 - e^{\frac{-x}{w_j}}) = 1 + \sum_{m=1}^{M-1} [(-1)^m \sum_{k=1}^{a_m} e^{-x(\sum_{j\in\Omega(k)} \frac{1}{w_j})}] \tag{5.5}$$

$$a_m = \binom{M-1}{m} \tag{5.6}$$

where a_m is the total number of possible combinations for selecting m SSs out of the $(M-1)$ SSs, and k represents the index of an arbitrary combination. Therefore, k is in the range of $[1, a_m]$. $\Omega(k)$ denotes the set of SSs corresponding to the combination index of k.

Thus, we have

$$\pi_i = 1 + \frac{1}{w_i} \sum_{m=1}^{M-1} [(-1)^m \sum_{k=1}^{a_m} 1/(\frac{1}{w_i} + \sum_{j\in\Omega(k)} \frac{1}{w_j})] \tag{5.7}$$

- Spectral Efficiency for SS_i

Spectral efficiency is defined as the amount of information bits transmitted over a unit bandwidth. The theoretical upper bound of spectral efficiency can be obtained based on Shannon's Formula. Given SS_i is selected for service when its weighted relative channel condition X_i equals x, the spectral efficiency achieved by SS_i is $log_2(1 + \frac{\overline{\gamma_i}}{w_i} x)$, where $\frac{\overline{\gamma_i}}{w_i} x$ is the corresponding instantaneous SNR of SS_i when its weighted relative channel condition is x. The probability that SS_i is selected for service with the weighted relative channel condition x is given by $l_i(x)(\prod\limits_{j=1,j\neq i}^{M}(\int\limits_{0}^{x} l_j(y)dy))$. Let ζ_i denote the spectral efficiency achieved by SS_i. Thus, the expectation of ζ_i is given by

$$\begin{aligned}
E[\zeta_i] &= \int\limits_{0}^{\infty}[log_2(1+\frac{\overline{\gamma_i}}{w_i}x)\cdot l_i(x)(\prod_{j=1,j\neq i}^{M}(\int\limits_{0}^{x} l_j(y)dy))]dx \\
&= \int\limits_{0}^{\infty}[log_2(1+\frac{\overline{\gamma_i}}{w_i}x)\cdot \frac{1}{w_i}e^{-\frac{x}{w_i}}(\prod_{j=1,j\neq i}^{M}(1-e^{-\frac{x}{w_j}}))]dx \\
&= \int\limits_{0}^{\infty}[log_2(1+\frac{\overline{\gamma_i}}{w_i}x)\cdot \frac{1}{w_i}e^{-\frac{x}{w_i}}[1+\sum_{m=1}^{M-1}(-1)^m\sum_{k=1}^{a_m}e^{-x(\sum\limits_{j\in\Omega(k)}\frac{1}{w_j})}]]dx \qquad (5.8)\\
&= \frac{1}{w_i}\int\limits_{0}^{\infty}log_2(1+\frac{\overline{\gamma_i}}{w_i}x)\cdot e^{-\frac{x}{w_i}}dx \\
&+ \frac{1}{w_i}\sum_{m=1}^{M-1}(-1)^m\sum_{k=1}^{a_m}\int\limits_{0}^{\infty}[log_2(1+\frac{\overline{\gamma_i}}{w_i}x)\cdot e^{-x(\frac{1}{w_i}+\sum\limits_{j\in\Omega(k)}\frac{1}{w_j})}]dx
\end{aligned}$$

where a_m and $\Omega(k)$ are defined as the same with that in (5.5), and we have

$$\begin{aligned}
&\int\limits_{0}^{\infty}(log_2(1+\frac{\overline{\gamma_i}}{w_i}x)\cdot e^{-x(\frac{1}{w_i}+\sum\limits_{j\in\Omega(k)}\frac{1}{w_j})})dx \\
&= \frac{e^{\frac{1+\sum\limits_{j\in\Omega(k)}\frac{w_i}{w_j}}{\overline{\gamma_i}}}}{\frac{\overline{\gamma_i}}{w_i}\cdot ln^2}\int\limits_{1}^{\infty}ln^y\cdot e^{-\frac{1+\sum\limits_{j\in\Omega(k)}\frac{w_i}{w_j}}{\overline{\gamma_i}}}dy \qquad (5.9)\\
&= \frac{e^{\frac{1+\sum\limits_{j\in\Omega(k)}\frac{w_i}{w_j}}{\overline{\gamma_i}}}}{(\frac{1}{w_i}+\sum\limits_{j\in\Omega(k)}\frac{1}{w_j})\cdot ln^2}\int\limits_{\frac{1+\sum\limits_{j\in\Omega(k)}\frac{w_i}{w_j}}{\overline{\gamma_i}}}^{\infty}e^{-t}t^{-1}dt
\end{aligned}$$

where $\int\limits_{\frac{1+\sum\limits_{j\in\Omega(k)}\frac{w_i}{w_j}}{\overline{\gamma_i}}}^{\infty}e^{-t}t^{-1}dt$ is the exponential integral function of first order for element $(\frac{1+\sum\limits_{j\in\Omega(k)}\frac{w_i}{w_j}}{\overline{\gamma_i}})$.

- Throughput for SS_i

Let Th_i^R denote the throughput achieved by SS_i based on the Rayleigh fading channel. Given a bandwidth W, Th_i^R is given as

$$Th_i^R = W \cdot E[\zeta_i] \tag{5.10}$$

- System Spectral Efficiency

From the system's point of view, the system spectral efficiency is the sum of achievable spectral efficiency of each SS, which is given by

$$\zeta = \sum_{i=1}^{M} E[\zeta_i] \tag{5.11}$$

where $E[\zeta_i]$ is the spectral efficiency achieved by SS_i.

- System Throughput

The system throughput is given by

$$Th^R = \sum_{i=1}^{M} Th_i^R \tag{5.12}$$

5.2.2 Performance Analysis with AMC

We further investigate the impact of the promising AMC technique, which has been specified in IEEE 802.16 standard. With AMC, the channel is characterized as an N-state Markov model described in Subsection 3.2.2.

- The Service Probability

The service probability only depends on the scheduling scheme and channel conditions. It does not impacted by whether the AMC is adopted or not. Therefore, the service probability with AMC can be derived by (5.7).

- Resource Utilization and Throughput for SS_i

In IEEE 802.16 networks with AMC, the resource utilization achieved at SS_i is defined as the information bits carried by an OFDM symbol.

Let τ_i denote the resource utilization achieved at SS_i. Its expectation is given by

$$\begin{aligned}
E[\tau_i] &= \sum_{n=1}^{N}[\int_{b_n}^{b_{n+1}}[I_n \cdot l_i(x)(\prod_{j=1,j\neq i}^{M}(\int_{0}^{x} l_j(y)dy))]dx] \\
&= \sum_{n=1}^{N}[\int_{b_n}^{b_{n+1}}(I_n \cdot \frac{1}{w_i}e^{-\frac{x}{w_i}}(\prod_{j=1,j\neq i}^{M}(1-e^{-\frac{x}{w_j}})))dx] \\
&= \sum_{n=1}^{N}[\int_{b_n}^{b_{n+1}}[\frac{I_n}{w_i}e^{-\frac{x}{w_i}}(1+\sum_{m=1}^{M-1}((-1)^m\sum_{k=1}^{a_m}e^{-x(\sum_{j\in\Omega(k)}\frac{1}{w_j})}))]dx] \\
&= \sum_{n=1}^{N} I_n(e^{-\frac{b_n}{w_i}} - e^{-\frac{b_{n+1}}{w_i}}) \\
&+ \sum_{n=1}^{N}\sum_{m=1}^{M-1}(-1)^m\sum_{k=1}^{a_m}(\frac{I_n}{1+\sum_{j\in\Omega(k)}\frac{w_i}{w_j}})(e^{-b_n(\frac{1}{w_i}+\sum_{j\in\Omega(k)}\frac{1}{w_j})} - e^{-b_{n+1}(\frac{1}{w_i}+\sum_{j\in\Omega(k)}\frac{1}{w_j})})
\end{aligned} \tag{5.13}$$

where b_n is the lower boundary of SNR for the channel state n, which is given in Table 3.1. Note that $b_{N+1} = \infty$. I_n is the information bit carried by an OFDM symbol corresponding to the channel state n, while a_m and $\Omega(k)$ are the same definition as that in (5.8).

Thus, the throughput achieved by SS_i considering AMC is given by

$$Th_i^A = E[\tau_i]/T_s \tag{5.14}$$

where T_s is the time duration of an OFDM symbol.

- System Resource Utilization and System Throughput

From the system's point of view, the system resource utilization is

$$E[\tau] = \sum_{i=1}^{M} E[\tau_i] \tag{5.15}$$

where M is the total number of SSs, and $E[\tau_i]$ is the resource utilization of SS_i.

Thus, the system throughput is

$$Th^A = \sum_{i=1}^{M} Th_i^A \tag{5.16}$$

where Th_i^A is the throughput of SS_i.

5.2.3 Fairness

Fairness is an important performance metric to evaluate the performance of a scheduling scheme. We use the Jain fairness index [93], a commonly adopted fairness index, to measure the fairness performance of a scheme, which is defined as

$$I = \frac{\left|\sum\limits_{k=1}^{M} h_k\right|^2}{M\sum\limits_{k=1}^{M}(h_k)^2} = \frac{1}{1+\frac{(\frac{1}{M}\sum\limits_{k=1}^{M} h_k^2)-(\frac{1}{M}\sum\limits_{k=1}^{M} h_k)^2}{(\frac{1}{M}\sum\limits_{k=1}^{M} h_k)^2}} = \frac{1}{1+COV^2} \tag{5.17}$$

where M is the total number of SSs, COV represents the coefficient of variation, and h_k is defined as

$$h_i = \begin{cases} \frac{r_i}{d_i} & if\ r_i < d_i \\ 1 & Otherwise \end{cases} \tag{5.18}$$

where d_i is the resource demanded by SS_i, and r_i is the resource allocated to SS_i.

For the proposed scheduling scheme, the resources available for BE traffic are allocated among all the SSs based on their intrinsic traffic patterns. Therefore, fairness is a metric to measure how close the network resource allocation of each SS to the pre-defined value. The ideal fairness performance is achieved when the coefficient of variation (COV) in (5.17) is 0. That is, fairness index I is 1. The value of COV can be anywhere between 0 to ∞. With the increase of COV, the fairness performance decreases. By manipulating the weight of each SS based on its traffic demand, the proposed scheme can flexibly adjust the resource allocation among SSs. Therefore, it is concluded that the proposed scheduling scheme can achieve satisfying fairness performance, which is verified by the simulation given in Section 5.3.

5.3 Simulation Results

Extensive simulations are conducted to demonstrate the efficiency and effectiveness of the proposed scheme and verify the analytical model in terms of important performance metrics, including the service probability, spectral efficiency, throughput, resource utilization, as well as fairness. Rayleigh fading channel model is adopted in the simulation, and the total number of SSs is 20. In order to study the impact of the channel conditions on the performance metrics, 20 SSs are divided into four groups with the average SNR (ASNR) of 10, 15, 20, and 25dB, respectively. Each group includes 5 SSs. Meanwhile, in order to evaluate the impact of employing different weights on the performance metrics, each SS in a common group (with the same ASNR) is assigned with different weights. We assume that the number of timeslots available for BE traffic at each DL sub-frame follows a uniform distribution in the range of [1,10], and consider the saturated case for BE traffic of each SS. Moreover, in order to evaluate the efficiency of the proposed scheme, the proportional fairness is adopted as a counterpart for the purpose of comparison. The simulation parameters are given in Table 5.2. We repeat the simulation 50 times with different random seeds and calculate the average value.

Figures 5.1-5.3 show the performance metrics based on the Rayleigh fading channel for both the proposed scheme (denoted as WPF) and the counterpart scheme (denoted as PF), where Sim and Ana represent the simulation results and the analytical results, respectively. Figure 5.1 shows the service probability of each SS. It is observed that the weight w_i is a key factor to affect the service probability of each SS for WPF. Among the SSs with the same ASNR, the larger weight an SS has, the larger service probability it can achieve. Meanwhile, it can be seen that the service probability of each SS is almost independent of its average channel condition. By averaging out the average channel

Table 5.2: Simulation Parameters.

MAC frame duration	2.5 *ms*
DL sub-frame duration	1.25 *ms*
UL sub-frame duration	1.25 *ms*
OFDM symbol duration	23.8 *us*
Bandwidth	10MHz
Index of SSs	1–5 , 6–10 ,11–15 ,16–20
Weight of SSs	1–5 , 1–5 , 1–5 , 1–5
Average SNR (dB)	10 , 15 , 20 , 25

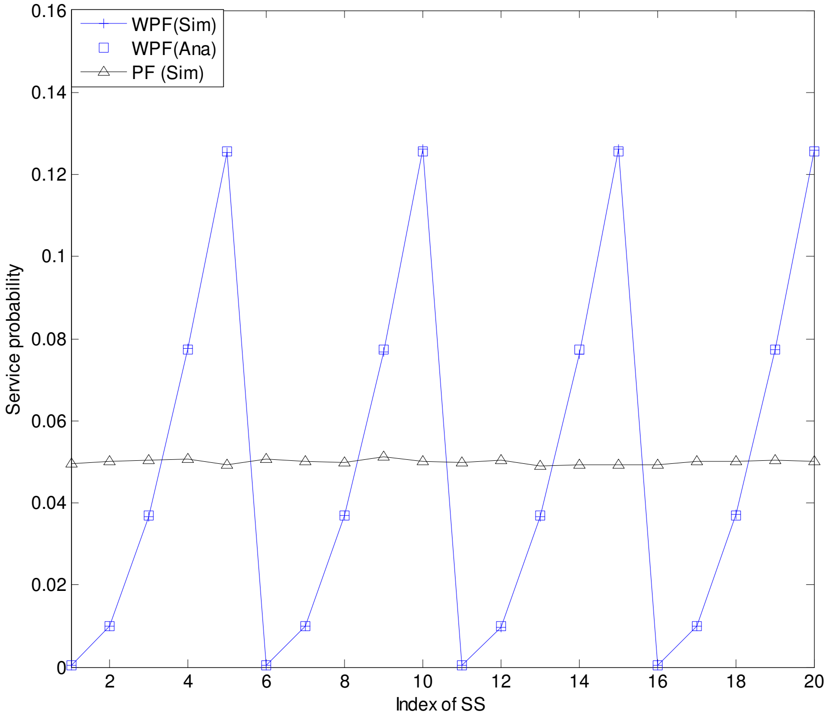


Figure 5.1: Service probability for each SS.

condition in the preference metric shown in (5.2), the effect of ASNR on the service probability of each SS is mitigated. This observation also supports the nice feature of

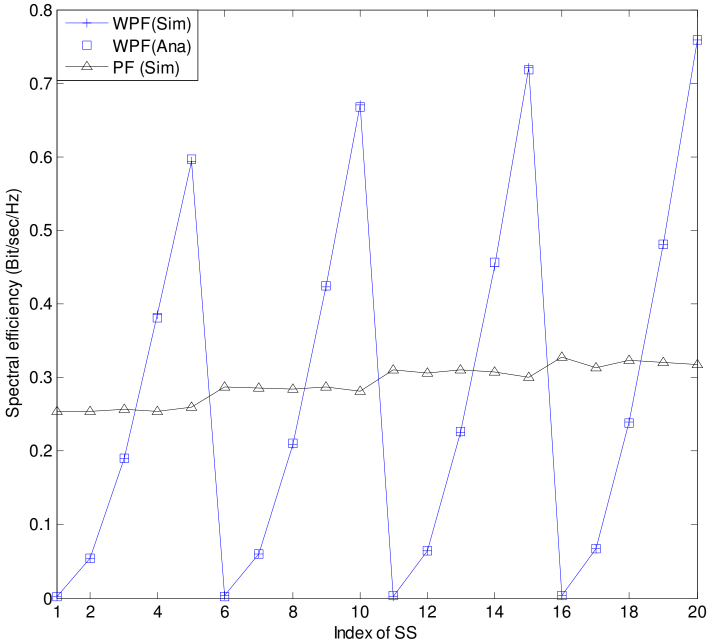


Figure 5.2: Spectral efficiency for each SS.

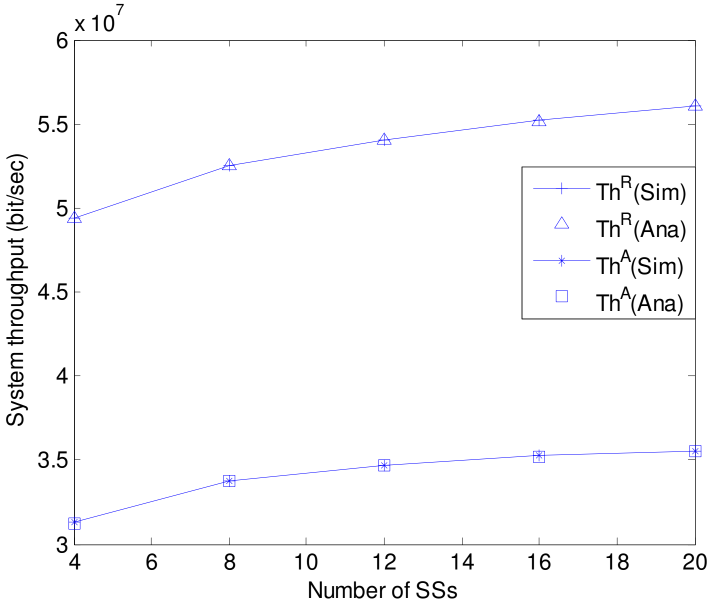


Figure 5.3: System throughput Versus the number of SSs.

the proposed scheme in terms of fairly scheduling all SSs since the scheduler will not be simply biased to the SSs due to their good average channel conditions. On the other hand, with PF, all SSs obtain similar service probabilities. Since PF aims to equally allocate the chance of transmission among all the SSs, it can not provide the service differentiation based on the different traffic demands.

Figure 5.2 shows the spectral efficiency achieved by each SS. It can be seen that with the PF, spectral efficiency varies lightly with the ASNR. With WPF, on the other hand, spectral efficiency of each SS is determined by both weight w_i and ASNR. With the same ASNR, the larger w_i an SS has, the higher spectral efficiency it can achieve. Meanwhile, with the same weight w_i, the better channel condition an SS has, the larger spectral efficiency it can achieve. SS_{20} and SS_5 have the same weights, but SS_{20} achieves a higher spectral efficiency than that of SS_5 due to its better average channel condition.

Figure 5.3 shows the system throughput versus the number of SSs with the Rayleigh channel(denoted as Th^R) and with the AMC (denoted as Th^A). It can be seen that the achieved system throughput increases with the increase of SSs. With a larger number of SSs, a larger channel-diversity gain can be exploited. Thus, the system throughput in terms of Th^R and Th^A increases accordingly.

Figures 5.4 – 5.5 show the resource utilization and achieved throughput of each SS with the consideration of AMC. It can be seen that both the resource utilization and throughput mainly depend on the weight w_i. SS_4 and SS_5 have the same ASNR, but the throughput and resource utilization of SS_5 is much larger than that of SS_4 due to its larger weight. Therefore, flexible resource allocation can be achieved by properly manipulating the weight of each SS.

Figure 5.6 shows the fairness index for both the WPF scheme and PF scheme. Two scenarios are investigated in the simulation: one is based on the Rayleigh fading channel (denoted as (Ray)), and the other considers the AMC (denoted as (AMC)). In the simulation, by taking the analytical result of throughput at each SS as its traffic demand, we evaluate how close the assigned system resources are to the corresponding

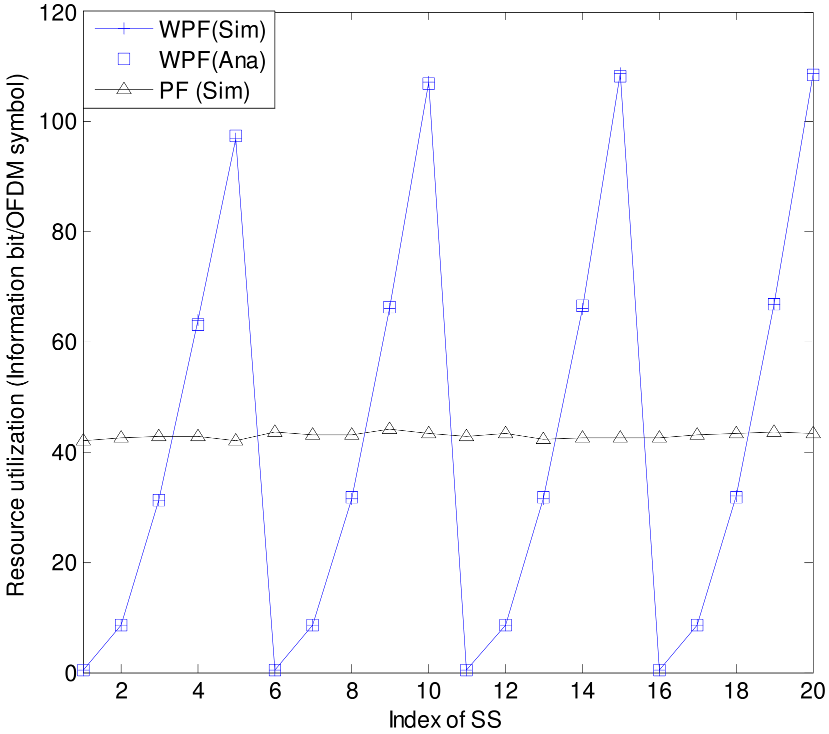


Figure 5.4: Resource utilization for each SS with discrete-rate AMC.

traffic demands. It is observed that the fairness performance of WPF outperforms PF. The fairness indices of the proposed scheme with both scenarios are very close to 1, which represents the ideal fairness performance. It also can be seen that the fairness performance of the proposed scheme is immune to the increasing number of SSs. On the contrary, for PF, the fairness performance exacerbates with the increase of SSs.

Figures 5.7 – 5.8 demonstrate system efficiency and resource utilization with the increase of SSs. It can be seen that both the system efficiency and system utilization increases with the increase of the number of SSs. With a larger number of SSs, a higher multi-user channel diversity gain can be exploited, which contributes to the overall system performance improvement in terms of system efficiency and resource utilization.

From Figures 5.1 – 5.8, it also can be seen that the simulation results and the

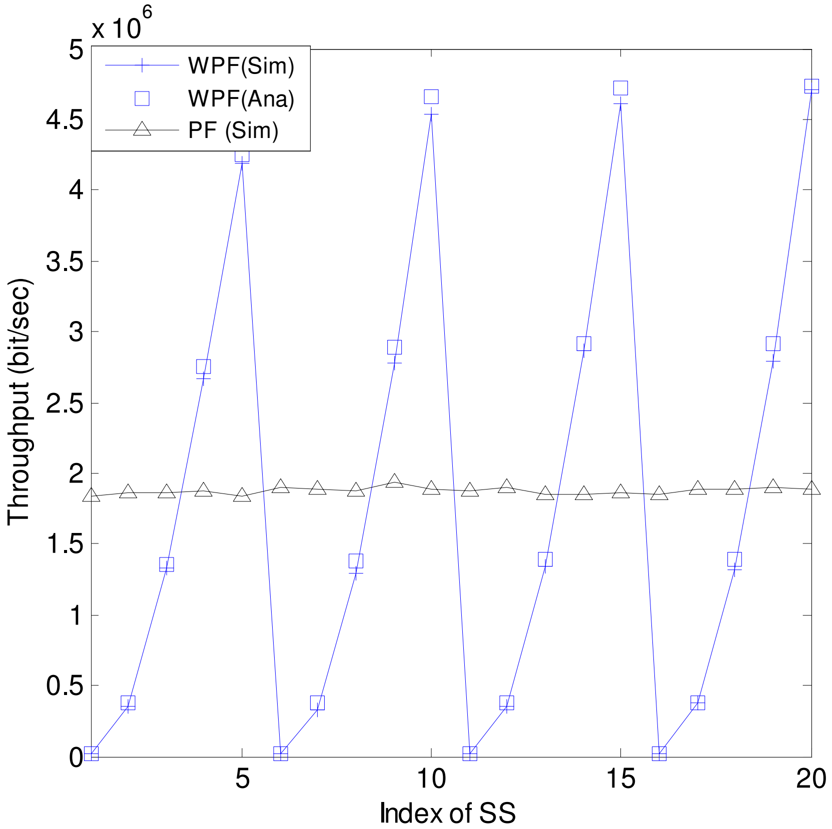


Figure 5.5: Throughput of each SS with discrete-rate AMC.

analysis results match very well, which verify the accuracy of the analytical model.

5.4 Discussion and Summary

In this chapter, we have proposed WPF scheduling scheme for BE traffic in IEEE 802.16 networks. The proposed scheme not only achieves the flexible resource allocation among heterogenous SSs, but also has a good fairness performance. An analytical model has been developed to investigate the performance of the proposed scheme in terms of throughput, spectral efficiency, and resource utilization. The analysis results can serve as a meaningful reference for the configuration of the weight for each SS under a specific design objective. Extensive simulations have been conducted to demonstrate

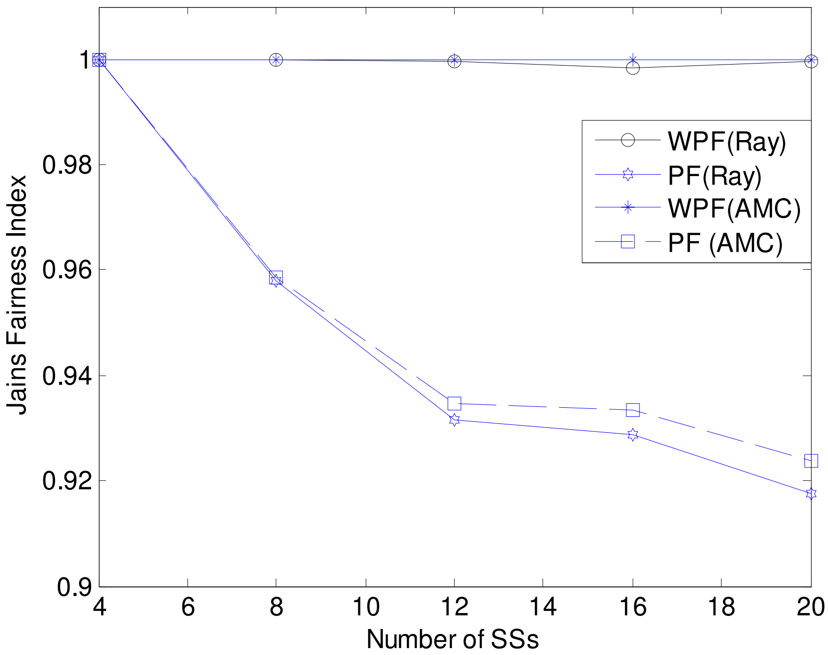


Figure 5.6: The fairness performance versus the number of SSs.

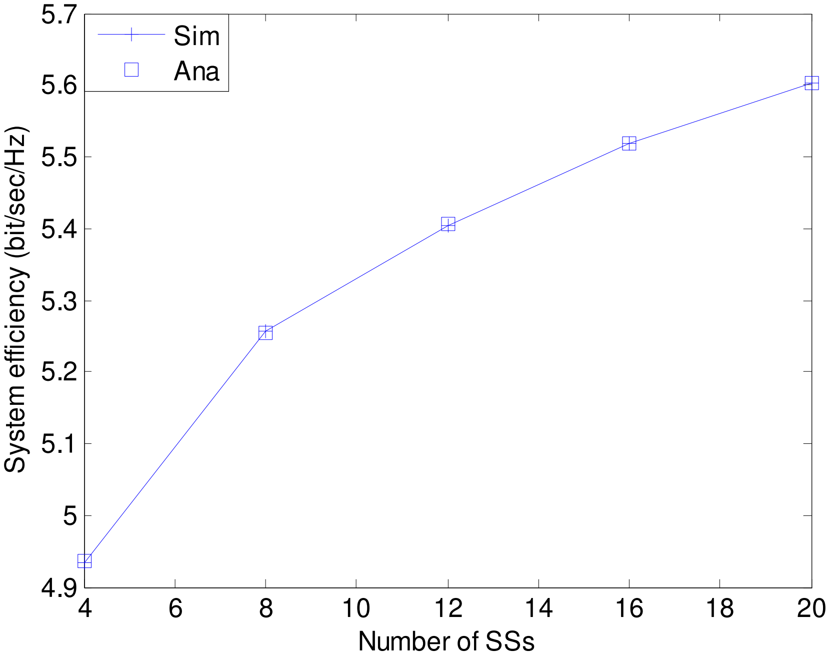


Figure 5.7: System efficiency versus the number of SSs.

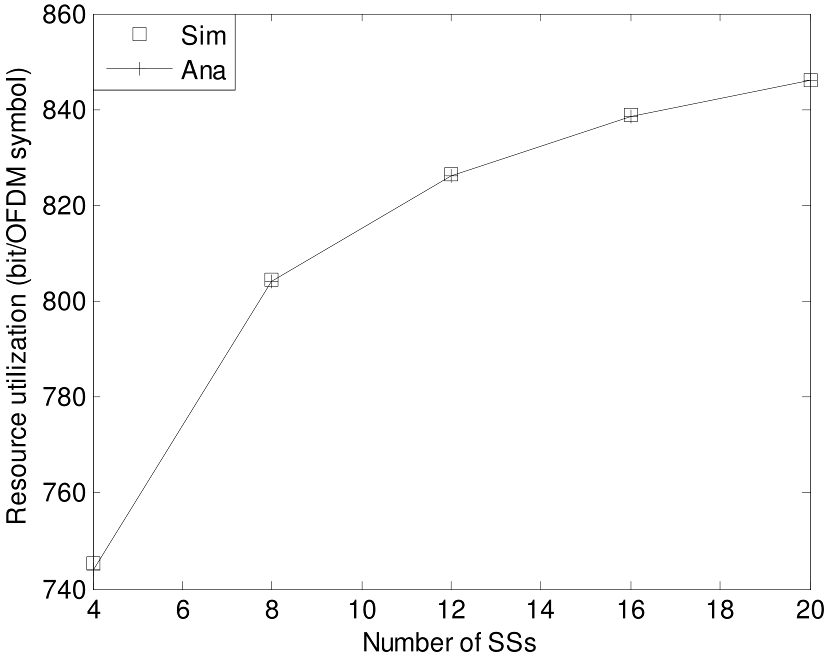


Figure 5.8: Resource utilization versus the number of SSs.

the efficiency of the proposed scheme and validate the analytical model.

The developed analytical model considers the Rayleigh fading channel in this chapter. However, it can be extended to other types of channel models with different probability distribution function for the received SNR. Furthermore, although this chapter is focused on IEEE 802.16 networks, the WPF scheduling scheme can be generally deployed in other types of networks composed of heterogenous types of users.

Chapter 6

Multicast Scheduling

With the increasing demand on multimedia applications, multimedia multicast services have been attracting great attentions from both academia and industry. Multimedia Broadcast Multicast Service (MBMS) has been standardized by the third generation partnership project (3GPP) and is currently under active investigation [59, 94]. Meanwhile, Broadband multimedia services, such as IPTV and mobile TV, are envisioned as major applications emerging in IEEE 802.16 WMANs, and expected to contribute immense market values to the service providers in next generation wireless networks [24, 25]. On the other hand, multicast transmission is an efficient way to provide services to multiple users simultaneously, by exploiting the broadcast nature of wireless communications. It is therefore critical to provide efficient multicast transmissions for supporting broadband multimedia services over IEEE 802.16 networks.

Most previous work on multicast transmissions is focused on efficient multicast routing protocols in the network layer [95–98] or effective error-control and recovery schemes in the transportation layer [99, 100]. Packet scheduling at the MAC layer plays a critical role in improving the resource utilization and providing QoS for multimedia multicast services. However, as discussed in Subsection 2.2.3, many previous studies on multicast scheduling aim at improving the group throughput at the expense of the transmission reliability of the group members in bad channel conditions. They are focused on mitigating the negative impact caused by the diverse channel conditions of multiple group members in an MGroup , rather than exploiting any potential advantage provided by the channel diversity among multiple group members in an MGroup. In a multicast scenario, multiple group members in an MGroup are independently located in the network. Since all group members in an MGroup need the same contents, the diverse channel conditions due to independent locations of group members can be exploited if some group members can help the other group members by transmitting the successfully received information. As a result, a more efficient and reliable multicast transmission should be provided, and the achieved throughput should be improved significantly.

Cooperative communication is a promising technology that can greatly improve the system performance by exploring the spatial diversity and cooperation among multiple users. Cooperative communication used for unicast transmissions has been extensively studied in the literature [101–106]. However, little work applies cooperative communication technique to multicast transmissions. Thus motivated, we propose a cooperative multicast scheduling scheme to efficiently exploit the spatial diversity among multiple users, based on a two-phase transmission model. In the first phase, the BS mulitcasts data at a high rate; and users in good channel conditions help relay the received data to the remaining users in the second phase. The proposed multicast cooperative scheme is different from unicast cooperative schemes in many aspects. First, the partner(s) or cooperator(s) in unicast cooperative transmissions are usually fixed, e.g., pre-placed relay stations, for protocol design and implementation simplicity. In the multicast scenario with all users in an Mgroup requesting the same data, basically any user with good channel conditions can forward the received data to the remaining users in the same group; and thus the cooperative transmitters are variable. Second, most previous studies in unicast cooperative transmission focus on the performance study in the

PHY layer, in terms of outage probability, bit error rate (BER), and optimal power allocation, etc. In a network scenario, users may have their own data to transmit besides forwarding the data of their partners. The key issue in this case is how to choose proper partners and efficiently coordinate the transmissions of relay data and the original data for each user. It is very difficult to analytically study the *network* performance of unicast cooperative schemes, and it becomes even harder when the number of cooperative transmitters is not fixed. In addition, different from unicast transmissions, multicast transmissions are inherently unreliable (due to no acknowledgement), and we need to carefully determine critical parameters for multi-user cooperation to assure high throughput for all users.

In this chapter, we propose an efficient multicast scheduling scheme for achieving the scalable and reliable multicast services. An analytical model is developed to evaluate the performance of the proposed scheme in terms of service probability, power consumption, and throughput of each group member and each MGroup. The proposed scheduling scheme can achieve high throughput, not only for all MGroups but also for each group member, by exploiting the spatial diversity among different group members and cooperative communication. In addition, the proposed channel-aware multicast group selection mechanism can guarantee fairness in terms of channel access by considering the normalized relative channel condition of each MGroup. Extensive simulations are conducted to demonstrate the effectiveness and efficiency of the proposed multicast scheduling scheme and the accuracy of the analytical model.

6.1 Cooperative Multicast Scheduling Scheme

In what follows, we investigate in detail the cooperative multicast scheduling scheme. In general, multiple SSs are classified into different MGroups according to their service requirements. For instance, for the IPTV service, an MGroup corresponds to a group of SSs requesting the same TV channel, whereas an SS could be a residential house or an office building, which may contain multiple end users watching different channels. Therefore, an SS may access multiple channels simultaneously and thus belongs to different MGroups.

For multicast scheduling scheme, the first key step is to select an appropriate MGroup for service at the beginning of each MAC frame; then the BS efficiently multicasts data to all group members in the selected MGroup, which are elaborated further as follows.

6.1.1 Multicast Group Selection

In this section, we introduce two approaches to select MGroups for service: the random MGroup selection and the channel-aware MGroup selection. The former one is basic and easy to implement, whereas the latter one can exploit multi-group channel diversity.

- **Random MGroup selection**

With random MGroup selection, the BS randomly selects an MGroup for service with a pre-defined probability. The probability for MGroup i to be selected in a MAC frame usually depends on the total number of MGroups. For instance, each group is served with the same probability $1/M$ for achieving a good fairness performance. The random MGroup selection scheme is easy to implement. In addition, flexible scheduling can also be achieved by setting different service probabilities to multiple MGroups according to their service demands.

- **Channel-aware MGroup selection**

To improve the throughput further, we propose a channel-aware MGroup selection. Different MGroups have different sets of group members distributed at different locations. Generally, group members experience different long-term channel conditions which depend on their geographical environments and the distances from the BS. On the other hand, due to small-scale fading, different group members may experience different instantaneous channel conditions at each frame, even if they have similar long-term channel conditions. In the proposed multicast scheduling scheme, to exploit the multi-group channel diversity gain, the selection of MGroups should consider the channel conditions on the group basis, rather than a single group member basis. If the selection of MGroups is based on the best channel condition among all members in an MGroup, ignoring the channel conditions of the remaining group members, the achievable group throughput may not be high if most of the other group members are experiencing bad channel conditions. If an MGroup is selected based on the overall channel conditions of the group members, it may lead to serious unfairness because MGroups which are close to the BS usually have good channel conditions, and thus are more likely to be scheduled and dominate the bandwidth consumption. By taking into account fairness while exploiting the multi-group channel diversity, we propose a criterion of MGoup selection based on the normalized relative channel condition, which is given by

$$i^* = \arg\max_i X_i \tag{6.1}$$

$$X_i = \frac{\sum\limits_{j\in G_i} \gamma_{i,j}/\overline{\gamma}_{i,j}}{N_i} \tag{6.2}$$

where X_i represents the normalized relative channel condition of MGroup i, G_i represents the set of all group members in MGroup i, N_i is the total number of group members in MGroup i, $\overline{\gamma}_{i,j}$ and $\gamma_{i,j}$ denote the average and instantaneous channel conditions of the jth group member in MGroup i, respectively. Based on (6.1), the BS selects MGroup i^*, which has the maximum value of the normalized relative channel condition, for service in each MAC frame.

To implement the channel-aware MGroup selection, BS should have the knowledge of the channel state information of each MGroup members. In IEEE 802.16 networks, the uplink channel quality indication channel (UL CQICH) is allocated for SSs to feedback channel state information. In addition, some SSs in IEEE 802.16 networks are stationary, such as office buildings or residence houses. This unique feature can largely decrease the frequency of channel state feedback since the channel conditions of these SSs are less fluctuant. Thus, the overhead of implementation is reduced. We define (6.2) as the preference metric of the channel-aware MGroup selection. Based on the channel state information of each MGroup member, a preference metric vector $\underline{X} = [X_1, X_2, \cdots, X_M]$ is maintained at the scheduler at the BS. At the beginning of each frame, $\underline{X}$ is updated according to (6.2). The BS selects the MGroup with the largest value of preference metric according to (6.1) and allocates the corresponding transmission burst to this MGroup.

In summary, by considering the channel conditions across multiple MGroups, the channel-aware MGroup selection exploits the multi-group channel diversity. On the other hand, by averaging out the long-term channel conditions and normalizing by the total number of MGroup members, the proposed scheduling scheme can achieve a good fairness performance as well. Note that the proposed channel-aware MGroup selection can achieve a better performance in terms of network throughput than random selection, at the cost of more overhead, such as signaling exchange, channel estimation and computation.

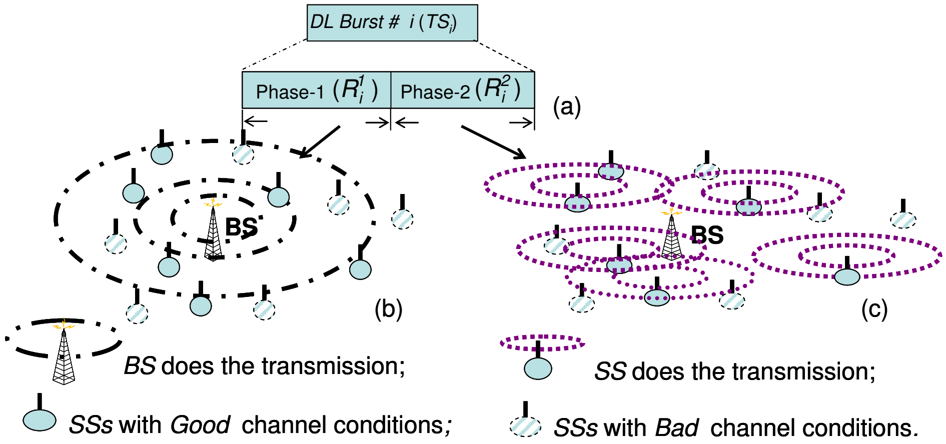

DL Burst # i (TS_i)
Phase-1 (R_i^1)
Phase-2 (R_i^2)
(a)
BS
(b)
BS
(c)
BS does the transmission;
SS does the transmission;
SSs with Good channel conditions;
SSs with Bad channel conditions.

Figure 6.1: The illustration of the cooperative multicast scheduling scheme.

6.1.2 Multicast Data

After an MGroup is selected, the next step is to efficiently multicast data to all group members in the selected MGroup. If the rate is set too high, some group members with bad channel conditions may not be able to successfully decode the data. On the contrary, if the rate is determined based on the group members with bad channel conditions, the wireless resources would be underutilized since the group members with good channel conditions can support a higher data rate. This dilemma is mainly caused by the diverse channel conditions of group members in the same MGroup. To exploit the spatial diversity gain of wireless channels, a two-phase transmission scheme is used to efficiently multicast data for the downlink transmissions, where a downlink burst is divided into two phases. For instance, MGroup i is selected for service in a frame and can access channel during the downlink burst TS_i. The time interval of TS_i is divided into two phases, as shown in Figure 6.1(a). In Phase I of time duration T_1, the BS multicasts data to all group members of MGroup i at a high data rate of R_i^1 such that only a certain portion of group members in MGroup i can successfully decode the data, as shown in Figure 6.1(b). Due to the high data rate, the remaining group members with bad channel conditions can not successfully decode the data in Phase I. Therefore, in the following Phase II, the cooperative communication is used to assure reliable transmissions of the remaining group members with bad channel conditions. Let S_i^g and S_i^b denote the set of group members that can and cannot successfully receive the data in Phase I, respectively. In Phase II of time duration T_2, all members in S_i^g transmit the received data to the members in S_i^b at the high rate of R_i^2, as shown in Figure 6.1(c). In this way, group members located in different locations form a virtual multiple input multiple output (MIMO) system, in which group members in S_i^g are transmitters and those in S_i^b are receivers. For a member in S_i^b, although the channel condition from the BS is relatively poor during this frame, the channel conditions between itself and some members in S_i^g may be good due to independent geographical locations of different group members. By exploiting the spatial diversity of wireless channels, group members in S_i^b are more likely to successfully receive the data in Phase II even at a high data rate. Therefore, the transmission rate for reliable multicast transmissions can be significantly improved.

One main advantage of the proposed scheme is that it can yield higher throughput, not only for group members with good channel conditions, but also for group members with bad channel conditions, by introducing cooperative communication and exploiting the multi-group channel diversity gain. Note that R_i^1 and R_i^2 are much higher than the conservative rate determined by the group member with the worst channel condition and the two-phase high rate transmission can outperform one phase conservative rate transmission [107]. Basically, it is conceptually possible to extend the two-phase transmissions to m-phase transmissions ($m > 2$). However, a large m involves more parameters and computation overhead, e.g., R_i^1, R_i^2, ... R_i^m, and may not always yield desirable network performance in terms of throughput and power consumption.

The transmission rates in Phases I and II (i.e., R_i^1 and R_i^2) are critical to the system performance. In the proposed scheme, R_i^1 and R_i^2 are determined based on the long-term channel conditions of all group members in MGroup i and the coverage ratio, C, which is defined as the percentage of group members that can support R_i^1. For instance, $C = 50\%$ means that the BS transmits at a rate of R_i^1 such that on average half of the group members in MGroup i can receive the data successfully, and R_i^2 is set in such a way that the remaining half of group members can successfully receive the data in Phase II based on their long term channel conditions. Meanwhile, the setting of the time durations of T_1 and T_2 satisfies $R_i^1 \cdot T_1 = R_i^2 \cdot T_2$ to assure all members in MGroup i can receive the same data. From the operation's point of view, the selection of R_i^1 and R_i^2 based on the long-term channel condition can lead to a easier implementation and less complexity since the BS does not need to reconfigure the transmission rate frequently.

6.2 Performance Analysis

In this Section, an analytical model is developed to investigate the network performance, including the service probability of each MGroup, the throughput of each group member, each MGroup and the whole network. The notations used in the rest of this section are listed in Table 6.1.

6.2.1 Performance Analysis Based on Channel Capacity

- **Service Probability for MGroup** i

Service probability is defined as the probability that an MGroup is selected for service at a frame when the system is stable. For the random MGroup selection, each MGroup is selected by a pre-defined probability. Thus, the steady-state service probability for MGroup i, π_i, is given as an operation parameter. For the channel-aware MGroup selection, according to (6.1), the MGroup with the largest normalized relative channel condition is selected. Define a random variable $Y_{i,j} = g_i(\gamma_{i,j}) = \frac{\gamma_{i,j}/\overline{\gamma}_{i,j}}{N_i}$, then $X_i = \sum\limits_{j \in G_i} Y_{i,j}$. Based on [92], the p.d.f. of $Y_{i,j}$ is given by

$$\phi(Y_{i,j} = y) = N_i e^{-N_i y} \tag{6.3}$$

where N_i is the total number of group members in MGroup i.

According to the relationship between X_i and $Y_{i,j}$, X_i has Gamma distribution, which is given by

$$X_i \sim Gamma(N_i, \frac{1}{N_i})$$

Table 6.1: Table of Notations

M	The total number of MGroups
G_i	The set of all members belonging to MGroup i
N_i	The total number of group members in MGroup i
G_i^g	The set of members in MGroup i that can successfully receive data in Phase I
G_i^b	The set of members in MGroup i that fail to receive data in Phase I
X_i	The normalized average channel condition of MGroup i
$SS_{i,j}$	The j-th group member in MGroup i
$\overline{\gamma}_{i,j}$	The average SNR of $SS_{i,j}$
$\gamma_{i,j}$	The instantaneous SNR of $SS_{i,j}$
$E_{i,j}^1$	The received signal power for $SS_{i,j}$ in Phase I
$E_{i,j}^2$	The received signal power for $SS_{i,j}$ in Phase II
$\overline{E}_{i,jB}$	The average received signal power for $SS_{i,j}$ from BS
$\overline{E}_{i,jk}$	The average received signal power for $SS_{i,j}$ from $SS_{i,k}$
N_0	The noise power
R_i^1	The transmission rate of the BS in Phase I for MGroup i
R_i^2	The transmission rate of each cooperative transmitter in Phase II for MGroup i
C	Coverage ratio used to set R_i^1
$Th_{i,j}^{CMS}$	The throughput of $SS_{i,j}$ for the proposed CMS scheme
$Th_{i,j}^{CON}$	The throughput of $SS_{i,j}$ for the multicast scheduling scheme *Conserve*
Th_i^{CMS}	The group throughput of MGroup i for the proposed CMS scheme
Th_i^{CON}	The group throughput of MGroup i for the multicast scheduling scheme *Conserve*

Thus, the service probability for MGroup i is given as

$$\begin{aligned}\pi_i &= Pr[X_i = max(X_1, X_2, ..., X_M)] \\ &= \int_0^\infty \left[h_i(X_i = x)(\prod_{j=1, j\neq i}^{M} H_j(X_j = x)) \right] dx \\ &= \int_0^\infty \left[\frac{{N_i}^{N_i}}{(N_i - 1)!} x^{N_i - 1} e^{-N_i x} \prod_{j=1, j\neq i}^{M} (1 - e^{-N_j x} \sum_{k=0}^{N_j - 1} \frac{(N_j x)^k}{k!}) \right] dx\end{aligned} \tag{6.4}$$

where the function h_i and H_j are the p.d.f. of X_i and the cumulative distribution function (C.D.F.) of X_j, respectively.

- **Throughput analysis based on channel capacity**

In the following, we study the throughput performance of group members with Rayleigh flat fading channels. Given a received SNR, the achievable channel capacity with a negligible error probability is $\log_2(1+SNR)$ for unit bandwidth [90] [91]. Therefore, given R_i^1 and R_i^2, the probability that a group member in MGroup i, $SS_{i,j}$, can successfully receive the data in Phase I is given by

$$Pr[E_{i,j}^1 \geq (2^{R_i^1}-1)N_0] = e^{-((2^{R_i^1}-1)N_0)/\overline{E}_{i,jB}}. \tag{6.5}$$

where $E_{i,j}^1$ is the received signal power of $SS_{i,j}$ in phase I, and N_0 is the noise power.

If $SS_{i,j}$ fails to receive the data in Phase I, it is still possible to successfully receive the data in Phase II. The received SNR of $SS_{i,j}$ in Phase II depends on the number of cooperative transmitters and the received signal power from each cooperative transmitter. Since all SSs in MGroup i except $SS_{i,j}$ are possible transmitters in Phase II, G_i^g could be any combination of these SSs, and we have

$$G_i^g \subseteq \{SS_{i,k}, k=1,2,...,N_i; k\neq j\} \tag{6.6}$$

Let $C_{i,j}$ be the set of all possible G_i^g for $SS_{i,j}$. The total number of all possible G_i^g is

$$|C_{i,j}| = \sum_{k=1}^{N_i-1}\binom{N_i-1}{k} = 2^{N_i-1}-1 \tag{6.7}$$

An example is given as follows. MGroup i is composed of four group members: $SS_{i,1}$, $SS_{i,2}$, $SS_{i,3}$, $SS_{i,4}$. For $SS_{i,1}$, the possible group members that can be its transmitters are $SS_{i,2}$, $SS_{i,3}$, and $SS_{i,4}$. Thus, the set of all possible G_i^g is $C_{i,1} = \{\{SS_{i,2}\},\{SS_{i,3}\},\{SS_{i,4}\},\{SS_{i,2},SS_{i,3}\},\{SS_{i,2},SS_{i,4}\},\{SS_{i,3},SS_{i,4}\},\{SS_{i,2},SS_{i,3},SS_{i,4}\}\}$, and the number of possible G_i^g for $SS_{i,1}$ is $|C_{i,j}| = \sum_{k=1}^{3}\binom{3}{k} = 7$. Given a MAC frame, the probability for any G_i^g to be the set of cooperative transmitters in Phase II is determined by their channel conditions. For instance, $G_i^g = \{SS_{i,2}, SS_{i,3}\}$ when only $SS_{i,2}$ and $SS_{i,3}$ can decode the data in Phase I. Thus, the probability of $G_i^g = \{SS_{i,2}, SS_{i,3}\}$ is given by

$$\begin{aligned} Pr\left(E_{i,2}^1 \geq (2^{R_i^1}-1)N_0\right) Pr\left(E_{i,3}^1 \geq (2^{R_i^1}-1)N_0\right) Pr\left(E_{i,4}^1 < (2^{R_i^1}-1)N_0\right) \\ = e^{-\frac{\left(2^{R_i^1}-1\right)N_0}{\overline{E}_{i,2B}}} e^{-\frac{\left(2^{R_i^1}-1\right)N_0}{\overline{E}_{i,3B}}} \left(1 - e^{-\frac{(2^{R_i^1}-1)N_0}{\overline{E}_{i,4B}}}\right) \end{aligned} \tag{6.8}$$

Let $E_{i,j}^2$ be the received signal power of $SS_{i,j}$ in Phase II. Thus, the probability that $SS_{i,j}$ can successfully receive the data in Phase II is given by

$$Pr\left(E_{i,j}^2 \geq (2^{R_i^2}-1)N_0\right) = \sum_{G_i^g \in C_{i,j}} Pr(G_i^g) Pr\left(E_{i,j}^2 \geq (2^{R_i^2}-1)N_0 | G_i^g\right) \tag{6.9}$$

where $Pr(G_i^g)$ denotes the probability that G_i^g is the set of transmitters in Phase II, $Pr(E_{i,j}^2 \geq (2^{R_i^2} - 1N_0)|G_i^g)$ is the probability that the received signal power of $SS_{i,j}$ in Phase II can support the sending rate R_i^2, given the G_i^g.

The received signal power in Phase II, $E_{i,j}^2$, is the sum of signal powers from all cooperative transmitters. For Rayleigh fading, the received signal power of $SS_{i,j}$ from the transmitter $SS_{i,k}, (SS_{i,k} \in G_i^g)$ has an exponential distribution. Thus, given G_i^g, $E_{i,j}^2$ is the sum of multiple random variables with independent exponential distributions. The close-form expression for the sum of squared Nakagami random variables is given in [108]. For Rayleigh fading channel, the p.d.f and C.D.F. of $E_{i,j}^2$ can be obtained by

$$f(E_{i,j}^2) = \sum_{SS_{i,k} \in G_i^g} \left[\frac{\overline{E}_{i,jk}}{\prod_{SS_{i,h} \in G_i^g} \overline{E}_{i,jh}} \left(\prod_{SS_{i,z} \in G_i^g; z \neq k} \left(\frac{1}{\overline{E}_{i,jz}} - \frac{1}{\overline{E}_{i,jk}}\right)^{-1} \right) \frac{1}{\overline{E}_{i,jk}} e^{-\frac{E_{i,j}^2}{\overline{E}_{i,jk}}} \right] \tag{6.10}$$

$$F(E_{i,j}^2) = \sum_{SS_{i,k} \in G_i^g} \left[\frac{\overline{E}_{i,jk}}{\prod_{SS_{i,h} \in G_i^g} \overline{E}_{i,jh}} \left(\prod_{SS_{i,z} \in G_i^g; z \neq k} \left(\frac{1}{\overline{E}_{i,jz}} - \frac{1}{\overline{E}_{i,jk}}\right)^{-1} \right) \left(1 - e^{-\frac{E_{i,j}^2}{\overline{E}_{i,jk}}}\right) \right] \tag{6.11}$$

where $E_{i,jk}$ is the received signal power of $SS_{i,j}$ from $SS_{i,k}$, and $\overline{E}_{i,jk}$ is the mean of $E_{i,jk}$.

Thus, the throughput achieved by the group member $SS_{i,j}$ is given as

$$\begin{aligned} Th_{i,j}^{CMS} &= \pi_i \Big[R_i^1 Pr(E_{i,j}^1 \geq (2^{R_i^1}-1)N_0) + R_i^2 Pr(E_{i,j}^1 < (2^{R_i^1}-1)N_0) Pr(E_{i,j}^2 \geq (2^{R_i^2}-1)N_0)\Big] \\ &= \pi_i \cdot [R_i^1 e^{-\frac{(2^{R_i^1}-1)N_0}{\overline{E}_{i,j}}} + R_i^2 (1 - e^{-\frac{(2^{R_i^1}-1)N_0}{\overline{E}_{i,j}}}) \\ & \quad \sum_{G_i^g \in C_{i,j}} [Pr(G_i^g)\ [1 - F(E_{i,j}^2 = (2^{R_i^2}-1)N_0 | G_i^g)]]] \qquad (6.12) \end{aligned}$$

where the function F is C.D.F. of received signal power of $SS_{i,j}$ in Phase II.

The group throughput achieved by MGroup i, which is defined as the summation of the throughput of all group members in MGroup i, is given by

$$Th_i^{CMS} = \sum_{j=1}^{N_i} Th_{i,j}^{CMS} \qquad (6.13)$$

The network throughput, which is defined as the summation of the group throughput of all MGroups in the network, is given by

$$Th^{CMS} = \sum_{i=1}^{M} Th_i^{CMS} \qquad (6.14)$$

We then study two extreme cases where all group members in MGroup i can or cannot support the sending rate R_i^1. The probability of these two cases are given in (6.15) and (6.16), respectively.

$$\prod_{SS_{i,j} \in G_i} Pr\left(E_{i,j}^1 \geq (2^{R_i^1}-1)N_0\right) = \prod_{SS_{i,j} \in G_i} e^{-\frac{(2^{R_i^1}-1)N_0}{\overline{E}_{i,jB}}} \qquad (6.15)$$

$$\prod_{SS_{i,j} \in G_i} Pr\left(E_{i,j}^1 < (2^{R_i^1}-1)N_0\right) = \prod_{SS_{i,j} \in G_i} \left(1 - e^{-\frac{(2^{R_i^1}-1)N_0}{\overline{E}_{i,jB}}}\right) \qquad (6.16)$$

6.2.2 Performance Analysis with AMC

In addition to the analysis based on the channel capacity, we further investigate the impact of the promising AMC technique at the PHY layer.

With the AMC technique, the received SNR is divided into several disjoint regions, based on a set of boundaries. Wireless channel is characterized as an N-state Markov model described in Subsection 3.2.2. Let b_n and I_n represent the lower boundaries of SNR and information bit carried by an OFDM symbol for the state n, respectively. b_n and I_n for different modulation and coding levels are given in Table 3.1.

The transmission rates corresponding to different modulation and coding levels are given by $T_n = I_n/T_s$, where T_s is the time duration of an OFDM symbol. Therefore, with AMC, the selection of R_i^1 and R_i^2 satisfies $R_i^1, R_i^2 \in \{T_n,\ n = 1, 2, \cdots, 7\}$. Given $R_i^1 = T_n$, and the probability that $SS_{i,j}$ can successfully receive the data in Phase I is

$$Pr(E_{i,j}^1 \geq b_n N_0) = e^{-(b_n N_0)/\overline{E}_{i,j} B} \tag{6.17}$$

Meanwhile, given $R_i^2 = T_m$, the probability that $SS_{i,j}$ can successfully receive the data in phase II is given by

$$Pr(E_{i,j}^2 \geq b_m N_0) = \sum_{G_i^g \in C_{i,j}} Pr(G_i^g) Pr(E_{i,j}^2 \geq b_m N_0 | G_i^g) \tag{6.18}$$

where b_n, b_m $(n, m = 1, 2, \cdots, 7)$ represent the the lower bound of SNR for the modulation and coding level n and m, respectively.

Therefore, with AMC technique, the throughput achieved by the group member $SS_{i,j}$ is given by

$$\begin{aligned} Th_{i,j}^{CMS} &= \pi_i \left[R_i^1\ Pr(E_{i,j}^1 \geq b_n N_0) + R_i^2\ Pr(E_{i,j}^1 < b_n N_0)\ Pr(E_{i,j}^2 \geq b_m N_0)\right] \\ &= \pi_i [R_i^1 e^{-\frac{(b_n N_0)}{\overline{E}_{i,j}}} + R_i^2 (1 - e^{-\frac{(b_n N_0)}{\overline{E}_{i,j}}}) \sum_{G_i^g \in C_{i,j}} [Pr(G_i^g)\ [1 - F(E_{i,j}^2 = b_m N_0 | G_i^g)]]] \end{aligned} \tag{6.19}$$

6.2.3 Performance Analysis for the Scheme *Conserve*

Multicast scheduling scheme *Conserve* is used as the counterpart in the simulation, where the BS selects a conservative transmission rate such that all group members in the selected MGroup can support this rate. Let φ_i denote the worst received SNR among all group members in MGroup i. It is give as

$$\varphi_i = min[\gamma_{i,1},\ \gamma_{i,2}, \cdots, \gamma_{i,N_i}] \tag{6.20}$$

where $\gamma_{i,j}(j = 1, 2, \cdots N_i)$ denote the received SNR of $SS_{i,j}$. For Rayleigh fading channel, $\gamma_{i,j}(j = 1, 2, \cdots N_i)$ has the exponential distribution and its p.d.f is given by

$$f(\gamma_{i,j}) = (1/\overline{\gamma}_{i,j})\ e^{\gamma_{i,j}/\overline{\gamma}_{i,j}} \tag{6.21}$$

where $\overline{\gamma}_{i,j}$ represents the average received SNR of $SS_{i,j}$ from the BS.

Thus, the p.d.f. of φ_i is given by

$$f(\varphi_i) = \left(\sum_{j=1}^{N_i} 1/\overline{\gamma}_{i,j}\right)\ e^{-\varphi_i\ \dot{(}\sum_{j=1}^{N_i} 1/\overline{\gamma}_{i,j})} \tag{6.22}$$

The throughput achieved by the group member $SS_{i,j}$ is given as

$$\begin{aligned} Th_{i,j}^{CON} &= \pi_i \int_{x=0}^{\infty} W log_2(1+\varphi_i) f(\varphi_i) d\varphi_i \\ &= \pi_i\ W\ \left(\sum_{j=1}^{N_i} 1/\overline{\gamma}_{i,j}\right) \int_{0}^{\infty} log_2(1+\varphi_i) e^{-\varphi_i(\sum_{j=1}^{N_i} 1/\overline{\gamma}_{i,j})} d\varphi_i \\ &= \pi_i\ W\ \frac{e^{(\sum_{j=1}^{N_i} 1/\overline{\gamma}_{i,j})}}{ln2} expint\left(\sum_{j=1}^{N_i} 1/\overline{\gamma}_{i,j}\right) \end{aligned} \tag{6.23}$$

where W is channel bandwidth, and $expint(x)$ is the exponential function of first order for element x, which is defined as $expint(x) = \int_{x}^{\infty} e^{-t}\frac{1}{t}dt$.

Thus, the achieved group throughput of MGroup i is given by

$$Th_i^{CON} = \sum_{j=1}^{N_i} Th_{i,j}^{CON} \tag{6.24}$$

6.3 Simulation Results

We compare the performance of the proposed multicast scheduling scheme (denoted as CMS) with the conservative scheme (denoted as $Conserve$) by extensive simulations with MATLAB. The IEEE 802.16 network is composed of one BS and 50 SSs. SSs are randomly distributed in the coverage area of the BS, which is a circle with a radius of 8 km. The group members in each MGroup is randomly selected from the 50 SSs. Other simulation parameters are listed in Table 6.2. We repeat the simulation 50 times with different random seeds and calculate the average value.

Table 6.2: Simulation Parameters

Transmission power of BS's	43 dBm
Transmission power of SS's	34.8 dBm
DL/UL sub-frame duration	1.25 ms/1.25 ms
OFDM symbol duration τ	23.8μs
Bandwidth	10 MHz
Noise figure	7 dB
Pass loss exponent	4.375
Close-in Reference distance	100 m
The total number of SSs in the system	50
The number of MGroups	10
The number of group members in each MGroup	20
Coverage ratio C	50%

The throughput performance is illustrated in Figures 6.2 – 6.3. The vertical axis is the achieved throughput normalized by the maximum value in the experiments. Figure 6.2 shows the throughput of each MGroup. Due to different geographical locations and channel conditions of each member in MGroups, the throughput varies in differ-

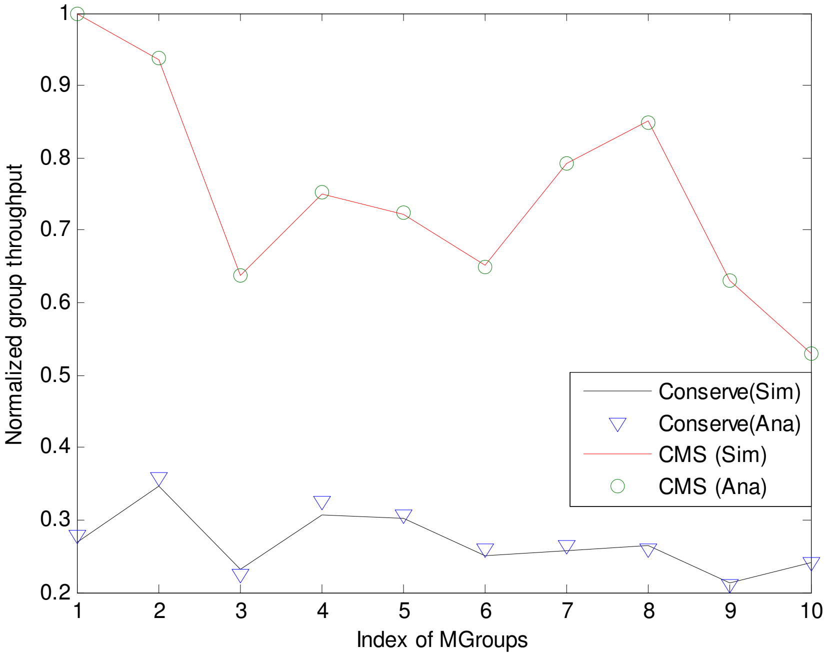


Figure 6.2: Group throughput of each MGroup.

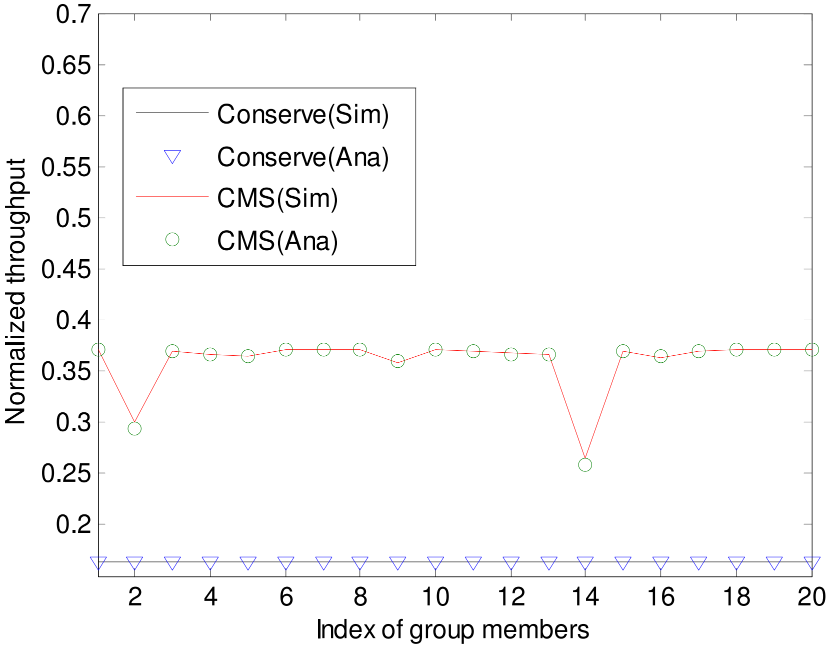


Figure 6.3: Throughput of each group members.

ent MGroups. It is observed that CMS outperforms $Conserve$ for all MGroups. The normalized throughput of each group member in an MGroup is shown in Figure 6.3. We observe that some isolated and faraway SSs achieve relatively lower throughput than other SSs. Generally, multimedia applications use scalable coding techniques, e.g., multi-layered video coding, and can tolerate some throughput fluctuations. For example, the group members with high throughput may receive both the base layer and enhancement layer information and thus can recover a high quality video, while other group members receive the base layer information and can only recover the basic quality video. With $Conserve$, all group members achieve the same throughput because they use the conservative transmission rate to ensure the successful transmissions of all SSs. By taking advantage of the spatial diversity and cooperation, the proposed CMS scheme significantly improves the throughput of all group members. Similar to Figure 6.3, the throughput performance based on the AMC is shown in Figure 6.4.

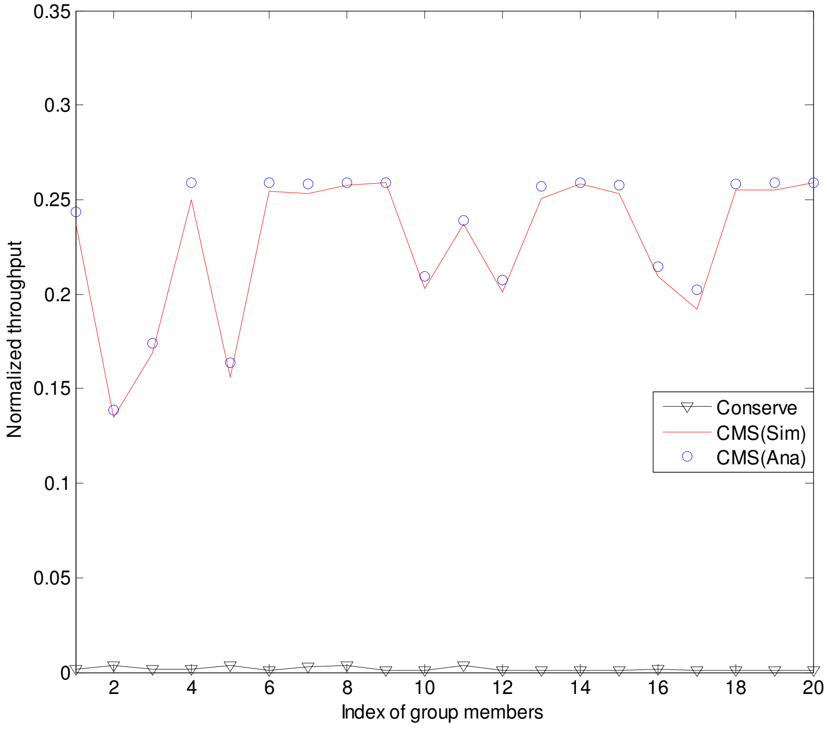


Figure 6.4: Throughput of each group members with AMC.

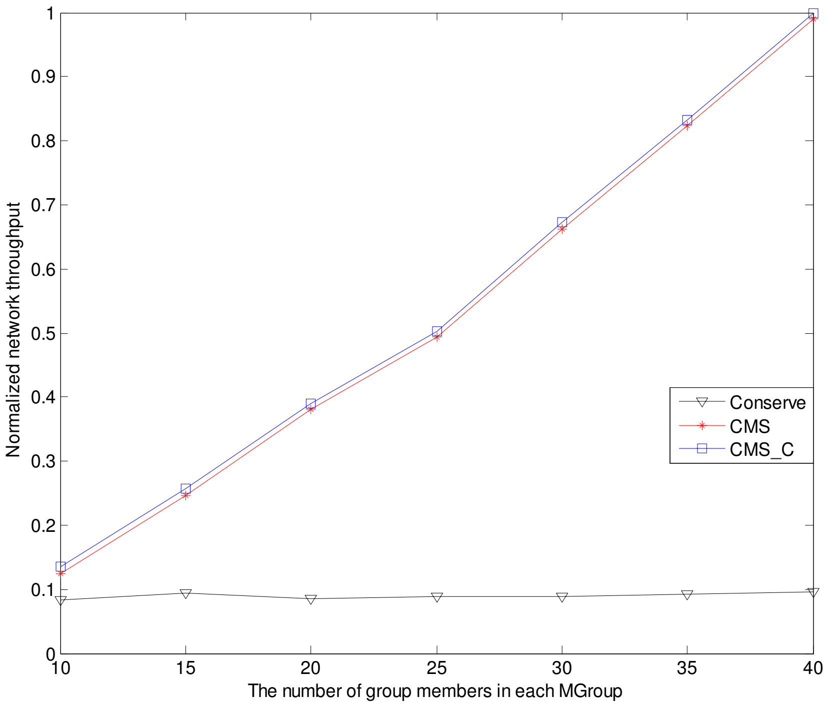


Figure 6.5: Network throughput versus the number of group members in each MGroup.

The performance of the proposed channel-aware MGroup selection scheme (denoted as CMS_C) is investigated and compared with CMS and $Conserve$ in Figure 6.5. The network throughputs of CMS and CMS_C are much higher than that of $Conserve$. The more number of group members in the network, the greater throughput improvement we can achieve. This is because higher diversity gain can be exploited among a larger number of group members. In addition, CMS_C outperforms CMS by taking advantage of the multi-group channel diversity with channel-aware MGroup selection. In Figure 6.6, we study the fairness performance of the proposed CMS_C in terms of the service probability of each MGroup. It is shown that each MGroup obtains almost the same service probability, which demontrates that the proposed channel-aware MGroup scheme can achieve good fairness performance in terms of channel access.

Besides throughput and fairness, power consumption is another important performance metric. The total power consumption in the network is defined as the power

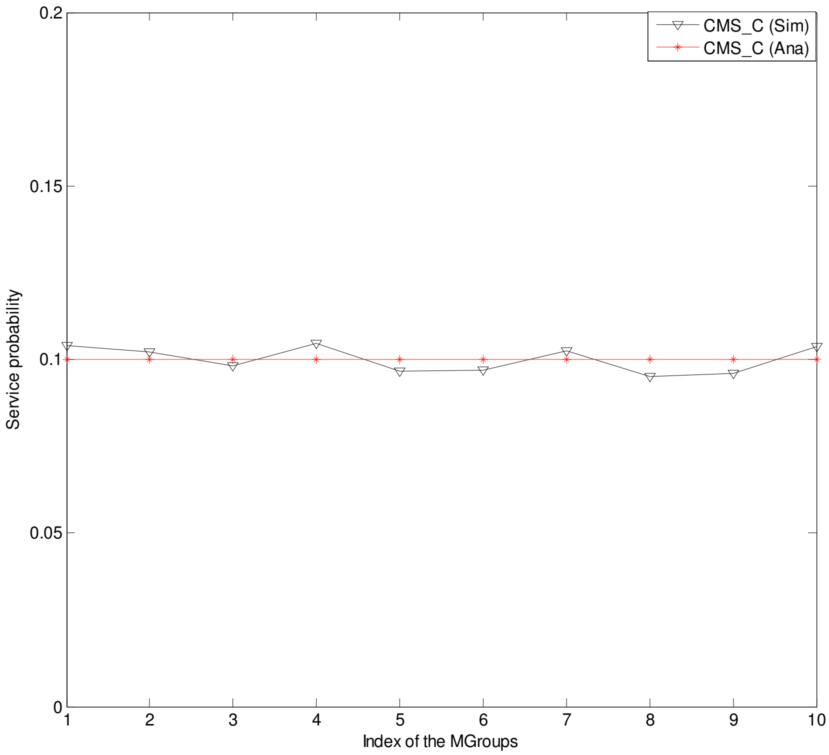


Figure 6.6: Service probability of each MGroup.

consumed by all transmitters, i.e., the BS and the involved transmitters in Phase II. As shown in Figure 6.7, the power consumption is a constant with *Conserve*, but it increases with the number of group members with *CMS* and *CMS_C*. This is because only the BS consumes power for downlink transmissions with *Conserve*. For the cooperative multicast scheduling scheme, although the BS does not transmit during Phase II, more SSs are likely to be involved in Phase II transmissions, resulting in a higher total power consumption. Comparing Figures 6.5 and Figure 6.7, we observe that *CMS* and *CMS_C* outperform *Converve* in terms of both throughput and power consumption when the number of group members in each MGroup is less than 15. With more group members, significant throughput improvement can be achieved with *CMS* and *CMS_C* at the expense of increased power consumption. As shown in Figures 6.5 and Figure 6.7, when the number of group member is 40, the power consumption of the

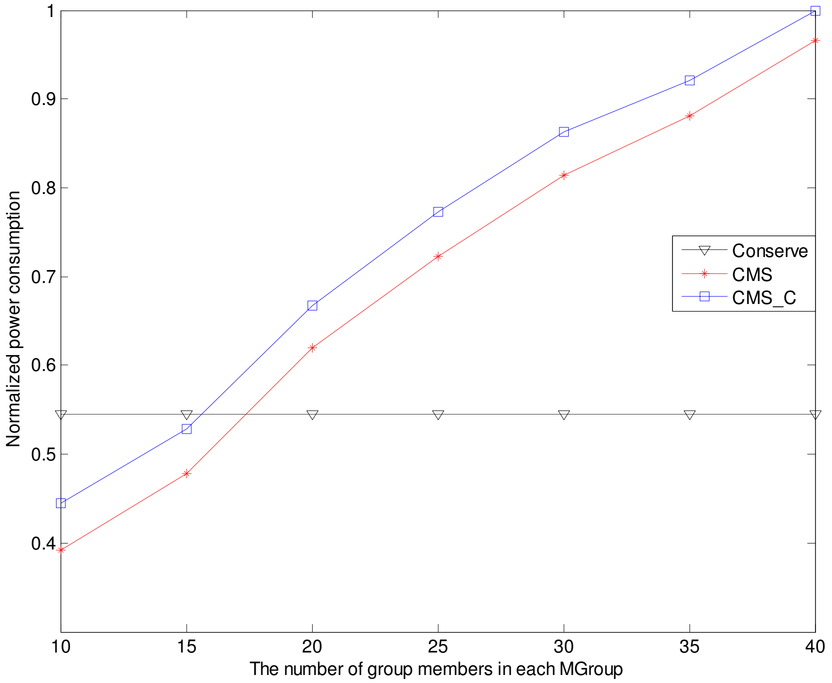


Figure 6.7: Power consumption versus the number of group members in each MGroup.

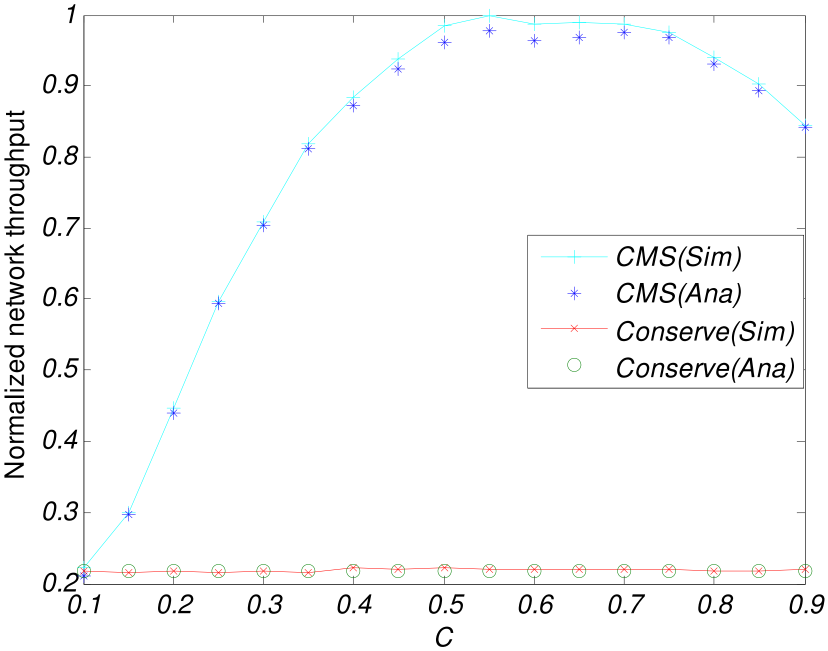


Figure 6.8: Network throughput versus the parameter C.

proposed CMS is around 1.7 times of that of $Conserve$, but the throughput of CMS is more than 10 times of that of $Conserve$.

Table 6.3 gives the analysis and simulation results for the probabilities of occurring two extreme cases, where case 1 and 2 represent that all group members in an MGroup can and cannot support the sending rate in phase I, respectively. It can be seen that the probabilities of these two extreme cases are less than 10^{-11} and 10^{-8}, respectively. Therefore, the impact of the extreme cases on the throughput is negligible.

Table 6.3: The probabilities of occurring the two extreme cases for each MGroup

Index of MGroups	1	2	3	4	5	6	7	8	9	10
Case 1 (Ana)(1e-11)	0.581	0	0.005	0.002	0.431	0	0.033	0.009	0.742	0.016
Case 2 (Ana)(1e-8)	0.586	0	0	0.002	0	0.119	0.758	0	0	0.031
Case 1 (Sim)	0	0	0	0	0	0	0	0	0	0
Case 2 (Sim)	0	0	0	0	0	0	0	0	0	0

We further study the impact of the coverage ratio C on the network throughput in Figure 6.8. The network throughput of the proposed scheme under different C values are always higher than that with $Conserve$. For the proposed CMS, the largest network throughput is achieved with $C = 0.55$, i.e., 55% SSs in an MGroup can forward the received data to the remaining SSs in Phase II. Simulation results validate the accuracy of our analysis.

6.4 Discussion and Summary

We have proposed a cooperative multicast scheduling scheme for achieving high throughput and good fairness to support broadband multimedia services in IEEE 802.16 networks. By using two-phase transmissions to exploit the spatial diversity gain in the multicast scenario, the proposed scheduling scheme can significantly improve the throughput not only for all MGroups, but also for each group member. In addition, the proposed channel-aware MGroup selection not only achieves a high network throughput by exploiting the multi-group channel diversity but also provides a good fairness among different MGroups in terms of channel access. Furthermore, an analytical model has been developed to evaluate the performance of the proposed scheme in terms of the service probability and the throughput, which provides a useful guideline for the system design and parameter setting, such as the number of TV channels supported by the network. Although the cooperative multicast scheduling scheme is proposed for IEEE 802.16 networks, it can also be extended to other wireless networks in general.

Chapter 7

Summary and Future Work

7.1 Summary of Research Contributions

In this thesis, we have studied the scheduling for two types of transmissions: unicast transmission and multicast transmission. For unicast transmission, nrtPS and BE service have been discussed. For multicast transmission, an efficient multicast scheduling scheme has been proposed to achieve a better performance in terms of throughput, fairness, and reliability. In specific,

- A simple yet efficient resource allocation and packet scheduling framework has been proposed for nrtPS applications in IEEE 802.16 networks, where two ARQ mechanisms at the MAC layer and AMC technique at the PHY layer are jointly considered. An analytical model has been developed to evaluate the performance of the proposed framework in terms of delivery delay, goodput, and resource utilization. The analytical results can provide meaningful guidelines to select appropriate parameters for satisfying the minimum throughput requirements and initiating a graceful compromise between the delivery delay and the resource utilization for nrtPS applications. Extensive simulations have been conducted to demonstrate the effectiveness and efficiency of the proposed framework and verify the accuracy of the analytical model;

- Considering the heterogeneity among SSs in terms of traffic load, we have proposed the weighted proportional fair scheduling scheme along with detailed implementation procedure for achieving flexible and fair scheduling and resource allocation for BE traffic in IEEE 802.16 networks. An analytical model has been developed to investigate some important performance metrics in terms of system spectral efficiency, resource utilization, throughput, and fairness, and quantify the impact of weights and channel conditions on them. The analytical model considers two scenarios. One is based on the capacity of Rayleigh fading channel, which provides the upper bounds on the spectral efficiency and throughput. The other considers AMC technique, which is specified in IEEE 802.16 standard. The analysis results can serve as a meaningful reference for the configuration of the weight of each SS under a specific design objective;
- Multicast transmission is an efficient way to providing one-to-many multimedia service. A cooperative multicast scheduling scheme has been proposed for achieving high throughput and good fairness to support broadband multimedia services in IEEE 802.16 networks. By considering the normalized relative channel condition of each MGroup, the proposed channel-aware MGroup selection can yield a good fairness among multiple MGroups in terms of channel access. Meanwhile, two-phase communication is introduced to efficiently multicast date to the selected MGroup by exploiting the spatical diversity gains in the multicast scenario. In Phased I, BS mulitcasts data at a high transmission rate, while in Phase II, the group members successfully received the data in Phase I help forward the data to other group members. By fairly selecting MGroups for service and exploiting the channel diversity of multiple group members, the proposed multicast scheduling scheme can improve the transmission rate and assure the transmission reliability of the group members with bad channel conditions, and hence result in significant throughput enhancement for both multicast groups and each individual group member; Furthermore, an analytical model on the proposed multicast scheduling scheme is developed to evaluate some important performance metrics, such

as the service probability of each MGroup, the power consumption, and the throughput of each group member and the whole network, which can provide a useful guideline for system design.

7.2 Further Work

The research work in this thesis focuses on the scheduling in IEEE 802.16 networks with PMP mode. Besides PMP mode, other kinds of network infrastructures such as multi-hop relay mode, mesh mode, and multi-cell scenario are of important and pose many challenging issues on the following aspects, which deserve further investigation.

- Relay is an efficient technique to enhance the system throughput and extend the coverage with a low cost. It has been attracted much attentions recently [109–111]. Mobile multihop relay (MMR) mode has been drafted in IEEE 802.16j developed by IEEE 802.16's Relay Task Group. The cooperative multicast scheduling scheme proposed and investigated in this thesis can achieve an enhancement of throughput not only for each MGroup but for each group member as well. However, due to the randomly distribution of group members in an MGroup, the fluctuation of throughout among SSs is observed, especially for some isolated and faraway SSs. For this case, relay is an efficient approach to decrease the fluctuation by enhancing the throughput of these isolated SSs. Therefore, how to further extend our proposed scheme considering the relay techniques to decrease the fluctuation and enhance the throughput is one of our further work.
- In recent years, the studies on the multi-cell wireless networks are attracting much attentions, where the resource allocation and scheduling are designed from the perspective of the whole networks. In [112, 113], network-wide scheduling schemes are proposed to achieve a better throughput and fairness in multi-cell networks. Channel allocation, power allocation, and load balance in multi-cell scenarios are addressed in [114–117] for improving the network-wide resource utilization, decrease the regional

congestion by alleviating the intra- and inter-cell interference and adaptively adjusting BS association. How to extend our work in the multi-cell wireless systems for achieving flexible and efficient resource allocation deserves further investigation.

- Cooperative communication at the BS-level is an efficient way to mitigate the interference in multi-cell scenarios, specially for users at the edge of the cell. With BS-level cooperation, the system resources can be allocated more efficiently by effectively controlling and alleviating the interference problem. Thus, the significant performance improvement can be achieved [118, 119]. BS association is one of the most important issues when addressing the BS-level cooperation. For different applications (e.g., unicast and multicast) and QoS requirements, the number of BSs associated to a single user or an MGroup and corresponding sending rates should be decided adaptively according to the channel conditions. Meanwhile, the power allocation and load balance among different cells are also interesting issues, which deserve further investigation in order to improve the resource utilization and the throughput of whole multi-cell network.

Bibliography

[1] A. Modarressi and S. Mohan, "Control and management in next-generation networks: challenges and opportunities," *IEEE Commun. Mag.*, vol. 38, no. 10, pp. 94–102, Oct. 2000.

[2] S. Parkvall, E. Englund, M. Lundevall, and J. Torsner, "Evolving 3G mobile systems: broadband and broadcast services in WCDMA," *IEEE Commun. Mag.*, vol. 44, no. 2, pp. 30–36, Feb. 2006.

[3] S. Lee, N. Park, C. Cho, and S. Ryu, "The wireless broadband (WiBro) system for broadband wireless Internet services," *IEEE Commun. Mag.*, vol. 44, no. 7, pp. 106–112, July 2006.

[4] C. Eklund, R.B. Marks, K.L, Stanwood, and S. Wang, "IEEE standard 802.16: a technical overview of the wirelessMAN air interface for broadband wireless access," *IEEE Commun. Mag.*, vol. 40, no. 6, pp. 98–107, June 2002.

[5] A. Ghosh, D.R. Wolter, J.G. Andrews, and R. Chen, "Broadband wireless access with WiMax/802.16: current performance benchmarks and future potential," *IEEE Commun. Mag.*, vol. 43, no. 2, pp. 129–136, Feb. 2005.

[6] R. Guerin and V. Peris, "Quality of service in packet networks: basic mechanisms and directions," *Computer Networks*, vol. 31, no. 3, pp. 169–179, Feb. 1999.

[7] V. Fineberg, "A practical architecture for implementing end-to-end QoS in an IP network," *IEEE Commun. Mag.*, vol. 40, no. 1, pp. 122–130, Jan 2002.

[8] S. Blake, D. Black, and M. Carlson, "Architecture for differentiated services," *IETF RFC 2475*, 1998.

[9] Y. Bernet, P. Ford, R. Yavatkar, F. Baker, et al., "A framework for integrated services operation over diffServ networks," *IETF RFC 2998*, Nov. 2000.

[10] W. Zhao, D. Olshefski, and H. Schulzrinne, "Internet quality of service: an overview," *Cloumbia Technical Report CUCS-003-00*, 2000.

[11] I. Stojmenovic and X. Lin, "Power-aware localized routing in wireless networks," *IEEE Trans. Parallel and Distributed System*, vol. 12, no. 11, pp. 1122–1133, Nov. 2001.

[12] V. Srinivasan, C.F. Chiasserini, P. Nuggehalli, and R.R. Rao, "Optimal rate allocation and traffic splits for energy efficient routing in ad hoc networks," in *Proc. IEEE INFOCOM*, vol. 2, New York, USA, June 2002, pp. 950–957.

[13] J. Gao and L. Zhang, "Load-balanced short-path routing in wireless networks," *IEEE Trans. Parallel and Distributed System*, vol. 17, no. 4, pp. 377–388, Apr. 2006.

[14] H. Wang, W. Li, and D.P. Agrawal, "Dynamic admission control and QoS for 802.16 wireless WAN," in *Wireless Telecommun. Sym.*, Pomona, USA, Apr. 2005, pp. 60–66.

[15] J. Ni, D. Tsang, S. Tatikonda, and B. Bensaou, "Threshold and reservation based call admission contorl policies for multiservice resource-sharing systems," in *Proc. IEEE INFOCOM*, vol. 2, Miami, USA, Mar. 2005, pp. 773–783.

[16] F. Hou, P.H. Ho, and X. Shen, "Performance analysis of a reservation based connection admission scheme in 802.16 networks," in *Proc. IEEE GLOBECOM*, San Francisco, USA, Nov. 2006, pp. 1–5.

[17] L. Cai, X. Shen, and J.W. Mark, "Congestion control for web-based multimedia playback applications," in *Proc. IEEE ICC*, vol. 1, Seattle, USA, May 2003, pp. 562–566.

[18] S. Pack, X. Shen, J.W. Mark, and L. Cai, "Throughput analysis of TCP friendly rate control in mobile hotspots," *IEEE Trans. Wireless Commun.*, vol. 7, no. 1, pp. 193–203, Jan. 2008.

[19] M. Andrews, K. Kumaran, A. Stolyar, P. Whiting, et al., "Providing quality of service over a shared wireless link," *IEEE Commun. Mag.*, vol. 39, no. 2, pp. 150–154, Feb. 2001.

[20] M. Mehrjoo, M. Dianati, X. Shen, and K. Naik, "Opportunistic fair scheduling for the downlink of IEEE 802.16 wireless metropolitan area networks," in *Proc. Qshine*, Waterloo, Ontario, Canada, Aug. 2006.

[21] C. Cicconetti, L. Lenzini, E. Mingozzi, and C. Eklund, "Quality of service support in IEEE 802.16 networks," *IEEE Network*, vol. 20, no. 2, pp. 50–55, Mar. 2006.

[22] A. Jalali, R. Padovani, and R. Pankaj, "Data throughput of CDMA-HDR a high efficiency-high data rate personal communication wireless system," in *Proc. IEEE VTC*, vol. 3, Tokyo, Japan, May 2000, pp. 1854–1858.

[23] H.J. Kushner, and P.A. Whiting, "Asymptotic properties of proportional-fair sharing algorithm," in *Proc. Allerton Conf. Commun. Control and Computing*, Illinois, USA, Oct. 2002, pp. 1051–1059.

[24] T. Jiang, W. Xiang, H. Chen, and Q. Ni, "Multicast broadcast services support in OFDMA-based WiMAX systems," *IEEE Commun. Mag.*, vol. 45, no. 8, pp. 78–86, Aug. 2007.

[25] J. She, F. Hou, P.H. Ho, and L.L. Xie, "IPTV over WiMAX: key success factors, challenges, solutions," *IEEE Commun. Mag.*, vol. 45, no. 8, pp. 87–93, Aug. 2007.

[26] G. Nair, J. Chou, T. Madejski, D. Putzolu, and J. Sydir, "IEEE 802.16 medium access control and service provisioning," *Intel Technology Journal*, vol. 08, no. 3, pp. 213–228, Aug. 2004.

[27] E. Agis, H. Mitchel, S. Ovadia, S. Aissi, et al., "Global interoperable broadband wireless netowrks: extending WiMAX technology to mobility," *Intel Technology Journal*, vol. 8, no. 3, pp. 173–187, Aug. 2004.

[28] S. J. Vaughan-Nichols, "Achieving wireless broadband with WiMax," *Computer*, vol. 37, no. 6, pp. 10–13, Jun. 2004.

[29] A. Ghosh, D.R. Wolter, J.G. Andrews, and R. Chen, "Broadband wireless access with WiMax/802.16:current performance benchmarks and future potential," *IEEE Commun. Mag.*, vol. 43, no. 2, pp. 129–236, Feb. 2005.

[30] N. Loutfi, *WiMAX: Technology for Broadband Wireless Access.* John Wiley, 2007.

[31] A. Sayenko, O. Alanen, J. Karhula, and T. Hamalainen, "Ensuring the QoS requirement in 802.16 scheduling," in *Proc. ACM MSWIM*, Torremolinos, Spain, Oct. 2006, pp. 108–117.

[32] IEEE 802.16a^{TM}-2003, "IEEE standard for local and metropolitan access network part 16: air interface for fixed broadband wireless access systems - amendment 2: medium access control modifications and additional physical layer specifications for 2-11 GHz," Apr. 2003.

[33] IEEE 802.16a^{TM}-2004, "IEEE standard for local and metropolitan access network part 16: air interface for fixed broadband wireless access systems," June 2004.

[34] Y. Cao, V.O.K. Li, "Scheduling algorithms in broad-band wireless networks," *Proc. IEEE*, vol. 89, no. 1, pp. 76–87, Jan. 2001.

[35] S. Lu, V. Bharghavan, and R. Srikant, "Fair scheduling in wireless packet networks," *IEEE/ACM Trans. Networking*, vol. 7, no. 4, pp. 473–489, Aug. 1999.

[36] T. Ng, I. Stoica, and H. Zhang, "Packet fair queueing algorithms for wirless networks with location-dependent errors," in *Proc. IEEE INFOCOM*, vol. 3, San Francisco, USA, Mar. 1998, pp. 1103–1111.

[37] J. Gomez, A.T. Campbell, and H. Morikawa, "The havana framework for supporting application and channel dependent QoS in wireless networks," in *Proc. Int. Conf. on Network Protocols*, Toronto, Canada, Nov. 1999, pp. 235–244.

[38] W. Kitti, and G. Aura, "Packet scheduling for QoS support in IEEE 802.16 broadband wireless access systems," *Int. J. on Commun. Syst.*, vol. 16, no. 1, pp. 81–96, Feb. 2003.

[39] Q. Liu, S. Zhou, G.B. Giannakis, "Cross-layer scheduling with prescribed QoS guarantees in adaptive wireless networks," *IEEE J. Select. Areas Commun.*, vol. 23, no. 5, pp. 1056–1066, May 2005.

[40] A.L. Stolyar, and K. Ramanan, "Largest weighted delay first scheduling: large deviations and optimality," *Ann. Appl. Prob.*, vol. 11, no. 1, pp. 1–48, Feb. 2001.

[41] S. Shakkottai, and A.L. Stolyar, “Scheduling algorithms for a mixture of real-time and non-real-time data in HDR,” in *Proc. Int. Teletraffic Congress*, Brazil, Sep. 2001, pp. 793–804.

[42] Q. Liu, X. Wang, and G. B. Giannakis, “A cross-layer scheduling algorithm with QoS support in wireless networks,” *IEEE Trans. Veh. Technol.*, vol. 55, no. 3, pp. 839–847, May. 2006.

[43] A.K.F. Khattab and K.M.F. Elsayed, “Opportunistic scheduling of delay sensitive traffic in OFDMA-based wireless networks,” in *Proc. IEEE Int. Sym. on World of Wireless Mobile and Multimedia Netowrks*, Buffalo, USA, June 2006, pp. 279–288.

[44] X. Liu, E.K.P. Chong, and N. B. Shroff, “A framework for opportunistic scheduling in wireless networks,” *Int. J. Computer and Telecomm. Networking*, vol. 41, no. 4, pp. 451–474, Mar. 2003.

[45] G.A. Amoakoh, “Multiuser diversity based opportunistic scheduling for wireless data networks,” *IEEE Commun. Letters*, vol. 9, no. 7, pp. 670–672, July 2005.

[46] M. Dianati, X. Shen, and K. Naik, “Per-user throughput of opportunistic scheduling scheme over broadcast fading channels,” in *Proc. IEEE ICC*, vol. 11, Istanbul, Turkey, June 2006, pp. 5234–5239.

[47] X. Liu, E.K.P. Chong, and N.B. Shroff, “Optimal opportunistic scheduling in wireless networks,” in *Proc. IEEE VTC*, vol. 3, Orlando, USA, Oct. 2003, pp. 1417–1421.

[48] R. Casaquite, I.Y. Kong, M.H. Yoon, and W.J. Hwang, “Opportunistic scheduling with power control in ad hoc wireless networks,” in *Proc. Int. Conf. on Advanced Commun. Technol.*, vol. 1, Phoenix Park , Korea, Feb. 2006, pp. 719–724.

[49] Y. AI-Harthi, A. Tewfik, and M.S. Alouini, “Opportunistic scheduling with quantized feedback in wireless networks,” in *Proc. Int. Conf. on Information Technology*, vol. 2, Las Vegas, USA, Apr. 2005, pp. 716–722.

[50] A.H. Ali, V. Krishnamurthy, and V.C.M. Leung, “Optimal and approximate mobility-assisted opportunisic scheduling in cellular networks,” *IEEE Trans. Mobile Comput.*, vol. 6, no. 6, pp. 633–648, June 2007.

[51] D. Avidor, S. Mukherjee, J. Ling, and C. Papadias, "On some properties of the proportional fair scheduling policy," in *Proc. IEEE PIMRC*, vol. 2, Barcelona, Spain, Sep. 2004, pp. 853–858.

[52] J-G Choi, and S. Bahk, "Cell throughput analysis of the proportional fair scheduling policy," in *Proc. Networking*, Athens, Greece, May 2004, pp. 247–258.

[53] D. Yang, D. Shen, W. Shao, and V.O.K. Li, "Towards opportunistic fair scheduling in wireless networks," in *Proc. IEEE ICC*, vol. 11, Istanbul, Turkey, June 2006, pp. 5217–5221.

[54] D.I. Kim, "Selective relative best scheduling for best-effort downlink packet data," *IEEE Trans. Wireless Commun.*, vol. 5, no. 6, pp. 1254–1259, June 2006.

[55] F. Berggren, and R. Jantti, "Asymptotically fair transmission scheduling over fading channel," *IEEE Trans. Wireless Commun.*, vol. 3, no. 1, pp. 326–336, Jan. 2004.

[56] J.H. Rhee, J.M. Holtzman, and D.K. Kim, "Performance analysis of the adaptive EXP/PF channel scheduler in an AMC/TDM system," *IEEE Commun. Letters*, vol. 8, no. 8, pp. 497–499, Aug. 2004.

[57] L. Xu, X. Shen, and J. W. Mark, "Dynamic fair scheduling with QoS constraints in multimedia wideband CDMA cellular networks," *IEEE Trans. Wireless Commun.*, vol. 3, no. 1, pp. 60–73, Jan. 2004.

[58] D. Park, H. Seo, H. Kwon, and B.G. Lee, "Wireless packet scheduling based on the cumulative distribution function of user transmission rates," *IEEE Trans. Commun.*, vol. 53, no. 11, pp. 1919–1929, Nov. 2005.

[59] 3GPP TR 25.992 V7.0.0, "Multimedia braodcast/multicast service (MBMS); UTRAN/GERAN requirement," $http://www.3gpp.org/ftp/Specs/archive/25-series/25.992/$, 2007.

[60] P. Agashe, R. Rezaiifar, and P. Bender, "CDMA2000 high rate broadcast packet data air interface design," *IEEE Commun. Mag.*, vol. 42, no. 2, pp. 83–89, Feb. 2004.

[61] P. Eusebio, and A. Correia, "Two QoS regions packet scheduling for multimedia broadcast multicast services," in *Proc. Int. Conf. on 3G and Beyond*, vol. 1, London, UK, Nov. 2005, pp. 34–38.

[62] C. Koh, and Y. Kim, "A proportional fair scheduing for multicast services in wireless cellular networks," in *Proc. IEEE VTC*, Montreal, Canada, Sep. 2006, pp. 1–5.

[63] H. Won, H. Cai, D.Y. Eun, K. Guo, et. al., "Multicast scheduling in cellular data networks," in *Proc. IEEE INFOCOM*, Anchorage, USA, May 2007, pp. 1172–1180.

[64] P. Chaporkar, S. Sarkar, "Wireless multicast: theory and approaches," *IEEE Trans. Information Theory*, vol. 51, no. 6, pp. 1951–1972, June 2005.

[65] W. Ge, J. Zhang, and X. Shen, "A cross-layer design approach to multicast in wireless networks," *IEEE Trans. Wireless Commun.*, vol. 6, no. 3, pp. 1063–1071, Mar. 2007.

[66] J.W. Mark and W. Zhuang, *Wireless Communications and Networking.* Prentice Hall, 2003.

[67] E. Gilbert, "Capacity of a burst-noise channel," *Bell Systems Technical Journal*, vol. 39, pp. 1253–1266, Sep. 1960.

[68] E. Elliott, "Estimates of error rates for codes on burst-noise channels," *Bell Systems Technical Journal*, vol. 42, pp. 1977–1997, Sep. 1963.

[69] M. Zorzi, R.R. Rao, and L.B. Milstein, "On the accuracy of a first-order markov model for data transmission on fading channels," in *Proc. Int. Conf. on Universal Personal Commun.*, Tokyo, Japan, Nov. 1995, pp. 211–215.

[70] D. Niyato, and E. Hossain, "A queuing-theoretic and optimization-based model for radio resource management in IEEE 802.16 broadband wireless networks," *IEEE Trans. Computers*, vol. 55, no. 11, pp. 1473–1488, Nov. 2006.

[71] C. Williamson, "Internet traffic measurement," *IEEE Int. Comput.*, vol. 5, no. 6, pp. 70–74, Nov. 2001.

[72] C. Fraleigh, S. Moon, B. Lyles, et. al., "Packet-level traffic measurements from the sprint IP backbone," *IEEE Network*, vol. 17, no. 6, pp. 6–16, Nov. 2003.

[73] W. -C. Wu, S. Vassiliadis, and T.Y. Chung, “Performance analysis of multi-channel ARQ protocols,” in *Proc. Midwest Symposium on Circuits and Systems*, vol. 2, Detroit, MI, USA, Aug 1993, pp. 1328–1331.

[74] J.G. Kim and M.M. Krunz, “Delay analysis of selective repeat ARQ for a markovian source over a wireless channel,” *IEEE Trans. Veh. Technol.*, vol. 49, no. 5, pp. 1968–1981, Sep. 2000.

[75] H. Shen, L. Cai, and X. Shen, “Performance analysis of TFRC over wireless links with truncated link level ARQ,” *IEEE Trans. Wireless Commun.*, vol. 5, no. 6, pp. 1479–1487, June 2006.

[76] F. Chiti, R. Fantacci, and R. Marchiani, “Performance analysis of a ARQ-SR protocol over a wireless packet network channel,” in *Proc. IEEE ICC*, vol. 11, Istanbul, Turkey, June 2006, pp. 5117–5122.

[77] F. Hou, P.H. Ho, and X. Shen, “A novel differentiated retransmission scheme for MPEG video streaming over wireless links,” *Int. J. Wireless and Mobile Computing*, vol. 1, no. 3/4, pp. 260–267, Feb. 2006.

[78] M. Zorzi, R.R. Rao, L.B. Milstein, “ARQ error control for fading mobile radio channels,” *IEEE Trans. Veh. Technol.*, vol. 46, no. 2, pp. 445–455, May 1997.

[79] W. Wang, Z. Guo, X. Shen, and C. Chen, “Performance analysis of ARQ scheme in IEEE 802.16,” in *Proc. IEEE GLOBECOM*, San Francisco, USA, Nov. 2006, pp. 1–5.

[80] R. Caceres, “Measurements of wide area internet traffic,” *Technical report UCB/CSD 89/550 computer science department, university of california, berkeley*, Dec. 1989.

[81] H. Wei, and R. Izmailov, “Channel-aware soft bandwidth guarantee scheduling for wireless packet access,” in *Proc. IEEE WCNC*, vol. 2, Atlanta, Georgia, Mar. 2004, pp. 1276–1281.

[82] V. Hassel, M.R. Hanssen, and G.E. ∅ien, “Spectral efficiency and fairness for opportunistic round robin scheduling,” in *Proc. IEEE ICC*, vol. 2, Istanbul, Turkey, June 2006, pp. 784–789.

[83] S.I. Hahm, H.J. Lee, and J.W. Lee, "A minimum-bandwidth guaranteed scheduling algorithm for data services in CDMA/HDR system," *Lecture Notes in Computer Science*, vol. 2713, pp. 757–763, June 2003.

[84] J.H. Lee, S.K. Kim, S.C. Lee, Y.J. Tcha, et.al., "Packet scheduling algorithm for non-real-time service with soft QoS requirement in mobile broadband wireless access system," in *Proc. IEEE ICC*, vol. 11, Istanbul, Turkey, June 2006, pp. 5240–5245.

[85] *Universal mobile telecommunications system (UMTS): selection procedures for the choice of radio transmission technologies of the UMTS (UMTS 30.03 version 3.2.0*, Std.

[86] H. Wang, and N. Moayeri, "Finite-state markov channel - a useful model for radio communication channels," *IEEE Trans. Veh. Technol.*, vol. 44, no. 1, pp. 163–171, Feb. 1995.

[87] F. Babich, and G. Lomardi, "A markov model for the mobile propagation channel," *IEEE Trans. Veh. Technol.*, vol. 49, no. 1, pp. 63–73, Jan. 2000.

[88] C. Pimentel, T. Falk, and L. Lisbôa, "Finite-state markov modeling of correlated rician-fading channels," *IEEE Trans. Veh. Technol.*, vol. 53, no. 5, pp. 1491–1501, Sep. 2004.

[89] Q. Zhang, and S. A. Kassam, "Finite-state markov model for rayleigh fading channels," *IEEE Trans. Commun.*, vol. 47, no. 11, pp. 1688–1692, Nov. 1999.

[90] M.S. Alouini, and A. J. Goldsmith, "Capacity of rayleigh fading channels under different adaptive transmission and diversity-combining techniques," *IEEE Trans. Veh. Technol.*, vol. 48, no. 4, pp. 1165–1181, July 1999.

[91] J. Li, A. Bose, and Y.Q. Zhao, "Rayleigh flat fading channels' capacity," in *Proc. Commun. Networks and Services Research Conf.*, Halifax, Canada, May 2005, pp. 214–217.

[92] P.J. Pahl, and R. Damrath, *Mathematical Foundations of Computational Engineering: A Handbook.* New York, Springer, 2001.

[93] R. Jain, D. Chiu, and W. Hawe, "A quantitative measure of fairness and discrimination for resource allocation in shared computer system," *Technical Report 301*, 1984.

[94] G. Zylomenos, V. Vogkas, and G. Thanos, "The multimedia broadcast/multicast service," *Wireless Commun. and Mobile Computing*, online, http://www3.interscience.wiley.com/cgi-bin/abstract/113390099/ABSTRACT, Oct. 2006.

[95] S. Lee, M. Gerla, and C. Chiang, "On-demand multicast routing rrotocol," in *Proc. IEEE WCNC*, New Orleans, USA, Sep. 1999, pp. 1298–1302.

[96] J.E. Wieselthier, G.D. Nguyen, and A. Ephermides, "On the construction of energy-efficient broadcast and multicast trees in wireless network," in *Proc. IEEE INFOCOM*, vol. 2, Tel-Aviv, Israel, Mar. 2000, pp. 585–594.

[97] X. Jia, D. Li, and F. Hung, "Multicast routing with minimum energy cost in ad hoc wireless networks," in *Proc. IEEE GLOBECOM*, vol. 5, Dallas, Texas, USA, Nov. 2004, pp. 2897–2901.

[98] M.X. Cheng, J. Sun, M. Min, Y. Li, and W. Wu, "Energy-efficient broadcast and multicast routing in multihop ad hoc wireless networks," *Wireless Commun. and Mobile Computing*, vol. 6, no. 2, pp. 213–223, Mar. 2006.

[99] J. Nonnenmacher, E. Biersack, and D. Towsley, "Parity-based loss recovery for reliable multicast transmission," *IEEE/ACM Trans. Networking*, vol. 6, no. 4, pp. 349–361, Aug. 1998.

[100] N. Nikaein, H. Labiod, and C. Bonnet, "MA-FEC: a QoS-based adaptive FEC for multicast communication in wireless networks," in *Proc. IEEE ICC*, vol. 2, New Orleans, USA, June 2000, pp. 954–958.

[101] A. Nosratinia, and A. Hedayat, "Cooperative communication in wireless networks," *IEEE Commun. Mag.*, vol. 42, no. 74-80, p. 10, Oct. 2004.

[102] P. Liu, Z. Tao, Z. Kiu, E. Erkip, and S. Panwar, "Cooperative wireless communications: a cross-layer approach," *IEEE Wireless Commun.*, vol. 13, no. 4, pp. 84–92, Aug. 2006.

[103] A. Sendonaris, E. Erkip, and B. Aazhang, "User cooperation diversity - part I: system description," *IEEE Trans. Wireless Commun.*, vol. 51, no. 11, pp. 1927–1938, Nov. 2003.

[104] A. Sendonaris, E. Erkip, and B. Aazhang, "User cooperation diversity - part II: implementation aspects and performance analysis," *IEEE Trans. Wireless Commun.*, vol. 51, no. 11, pp. 1939–1948, Nov. 2003.

[105] T.E. Hunter, S. Sanayei, A. Nosratinia, "Outage analysis of coded cooperation," *IEEE Trans. Information Theory*, vol. 52, no. 2, pp. 375–391, Feb. 2006.

[106] D.H. Woldegebreal, S. Valentin, and H. Karl, "Outage probability analysis of cooperative transmission protocols without and with network coding: inter-users based comparison," in *Proc. ACM MSWIM*, Chania, Greece, Oct. 2007, pp. 36–44.

[107] H. Ochiai, "Variable-rate two-phase collaborative communication protocols for wireless networks," *IEEE Trans. Information Theory*, vol. 52, no. 9, pp. 4299–4313, Sep. 2006.

[108] G.K. Karagiannidia, N.C. Dagias, and T.A. Tsiftsis, "Closed-form statistics for the sum of squared nakagami-m variates and its applications," *IEEE Trans. Commun.*, vol. 54, no. 8, pp. 1353–1359, Aug. 2006.

[109] Q. Wang, K. Xu, G. Takahara, and H. Hassanein, "Locally optimal relay node placement in heterogeneous wireless sensor networks," in *Proc. IEEE GLOBECOM*, vol. 6, St. Louis, USA, Nov. 2005, pp. 3549–3553.

[110] A. So and B. Liang, "Enhancing WLAN capacity by strategic placement of tetherless relay points," *IEEE Trans. Mobile Comput.*, vol. 6, no. 5, pp. 474–487, May 2007.

[111] B. Lin, P.H. Ho, L. Xue, and X. Shen, "Relay stateion placement in IEEE 802.16j dual-relay MMR networks," in *Proc. IEEE ICC*, Beijing, China, May 2008.

[112] J. Cho, J. Mo, and S. Chong, "Joint network-wide opportunistic scheduling and power control in multi-cell networks," in *Proc. IEEE Int. Sym. on World of Wireless Mobile and Multimedia Netowrks*, Helsinki, Finland, June 2007, pp. 1–12.

[113] T. Bu, L. Li, and R. Ramjee, "Generalized proportioal fair scheduling in third generation wireless data networks," in *Proc. IEEE INFOCOM*, Barcelona, Spain, Apr. 2006, pp. 1–12.

[114] H.K. Lim, J.G. Choi, and S. Bahk, “Utility-based downlink power allocation in multicell wireless packet networks,” in *Proc. IEEE GLOBECOM*, vol. 5, Dallas, USA, Nov. 2004, pp. 3321–3325.

[115] R. Rummler, Y.W. Chung, A.H. Aghvami, “Modeling and analysis of an efficient multicast mechanism for UMTS,” *IEEE Trans. Veh. Technol.*, vol. 54, no. 1, pp. 350–365, Jan. 2005.

[116] A. Sang, X. Wang, M. Mahihian, and R.D. Gitlin, “Coordinated load balancing, handoff/cell-site selection, and scheduling in multi-cell packet data systems,” in *Proc. ACM/IEEE MOBICOM*, Philadelphia, USA, Sep. 2004, pp. 302–314.

[117] N.H. Lee and S. Bahk, “Dynamic chanel allocation using the interference range in multi-cell downlink systems,” in *Proc. IEEE WCNC*, HongKong, China, Mar. 2007, pp. 1716–1721.

[118] H. Zhang and H Dai, “Cochannel interference mitigation and cooperative processing in downlink multicell multiuser MIMO networks,” *Eurasip J. on Wireless Commun. Networking*, vol. 4th quarter, no. 2, pp. 222–235, 2004.

[119] S. Shamai, Q. Somekh, O. Simeone, and A. Sanderovich, “Cooperative multi-cell networks: impact of limited-capacity backhaul and inter-users links,” in *Proc. Joint Workshop on Coding and Commun.*, Durnstein, Austria, Oct. 2007.

List of Abbreviations

3GPP	The Third Generation Partnership Project
ARQ	Automatic Repeat reQuest
ACK	Acknowledgement
AMC	Adaptive Modulation and Coding
ASNR	Average Signal-to-Noise Ratio
BE	Best Effort
BER	Bit Error Rate
BPSK	Binary Pulse Shift Keying
BS	Base Station
BWA	Broadband Wireless Access
CAC	Connection Admission Control
CARR	Channel-Aware Round Robin
CARQ	Cumulative ARQ
CBR	Constant Bit Rate
CDMA	Code Division Multiple Access
CM	Cable Modem
CMS	Cooperative Multicast Scheduling
CPS	Common Part Sub-layer
CS	Convergency Sub-layer

CSIR	Contention Slot for Initial Ranging
CSBR	Contention Slot for Bandwidth Request
DiffServ	Differentiated Service
DL sub-frame	DownLink sub-frame
DSL	Digital Subscriber Line
ertPS	extended real-time Polling Service
ETSI	European Telecommunications Standard Institute
FCC	Federal Communications Commission
FCH	Frame Control Header
FCH	Frame Control Header
FDD	Frequency Division Duplexing
FEC	Forward Error Control
FSMC	Finite State Markov Channel
FSN	Fragment Sequence Number
FTP	File Transfer Protocol
HFC	Hybrid Fiber Coax
HOL	Head of Line
IEEE	Institute of Electrical and Electronics Engineering
IETF	Internet Engineering Task Force
IntServ	Integrated Service
IPTV	Internet Protocol TeleVision
LOS	Line of Sight
LWDF	Largest Weighted Delay First
MAC	Media Access Control
MAN	Metropolitan Area Network

MBMS	Multimedia Broadcast Multicast Service
MIMO	Multiple Input Multiple Output
MGroup	Multicast Group
MMR	Mobile Multi-hop Reply
NLOS	Non Line of Sight
nrtPS	Non-real-time Polling Service
OFDM	Orthogonal Frequency Division Multiplex
OFDMA	Orthogonal Frequency Division Multiple Access
PCM	Pulse Coded Modulation
PDU	Protocol Data Unit
PF	Proportional Fairness
PHY	Physical
PMP	Point-to-Multi-Point
QAM	Quadrature Amplitude Modulation
QoS	Quality of Service
QPSK	Quadrature Phase Shift Keying
rtPS	Real-time Polling Service
SARQ	Selective ARQ
SDU	Service Data Unit
SNR	Signal-to-Noise Ratio
SS	Subscriber Station
SP	Strict Priority
TCP	Transmission Control Protocol
TDMA	Time Division Multiple Access
TDD	Time Division Duplexing

TDM	Time Division Multiplexing
UGS	Unsolicited Grant Service
UL sub-frame	Uplink sub-frame
UWB	Ultra-WideBand
UMTS	Universal Mobile Telecommunications Systems
UL CQICH	Uplink Channel Quality Indication Channel
VoIP	Voice over Internet Protocol
WiMAX	Worldwide Interoperability Microwave Access
Wi-Fi	Wireless Fidelity
WLAN	Wireless Local Area Network
WMAN	Wireless Metropolitan Area Networks
WPAN	Wireless Personal Area Networks
WPF	Weighted Proportional Fairness

List of Notations

B	The size of a PDU
T	The time duration of a MAC frame
L	The number of PDUs transmitted by the tagged queue in each transmission opportunities
h	The number of selected SSs for service in an DL transmission
p	Error probability of transmitting a PDU
σ_{S_n}	The probability of obtaining the service given the channel state n
m	The inter-service time
μ	The number of PUDs successfully lunched during each transmission opportunity with cumulative ARQ
η	The number of PUDs successfully lunched during each transmission opportunity with selective ARQ
G^C	The achieved goodput with cumulative ARQ
G^S	The achieved goodput with selective ARQ
D_P^C	The delivery delay of a PDU with the cumulative ARQ
D_P^S	The delivery delay of a PDU with the selective ARQ
D_S^C	The delivery delay of an SDU with the cumulative ARQ
D_S^S	The delivery delay of an SDU with the selective ARQ
τ	System resource utilization
$E[N_P]$	The expected number of transmission opportunities for successfully transmitting a PDU
$E[N_S]$	The expected number of transmission opportunities for successfully

	transmitting an SDU
M	The total number of SSs in the network
$\bar{\gamma}_i$	The average SNR of SS_i
γ_i	The instantaneous SNR of SS_i
w_i	The weight of SS_i
D_i	BE traffic demand of SS_i
X_i	The value of preference metric of SS_i
π_i	Service probability of SS_i
ψ	Index of SS selected for service during an DL frame
ς_i	Spectral efficiency for SS_i
ς	System spectral efficiency
τ_i	Resource utilization for SS_i
W	Channel bandwdith
T_s	The time duration of an OFDM symbol
Th_i^R	The throughput of SS_i with the Rayleigh fading channel
Th_i^A	The throughput of SS_i with the adaptive modulation and coding
Th^R	The system throughput of SS_i with the Rayleigh fading channel
Th^A	The system throughput of SS_i with the adaptive modulation and coding
M	The total number of MGroups
G_i	The set of all members belonging to MGroup i
N_i	The total number of group members in MGroup i
G_i^g	A set of members in MGroup i that can successfully receive data in Phase I
G_i^b	A set of members in MGroup i that fail to receive data in Phase I
X_i	The normalized average channel condition of MGroup i
$SS_{i,j}$	the j-th group member in MGroup i
$\bar{\gamma}_{i,j}$	The average SNR of $SS_{i,j}$
$\gamma_{i,j}$	The instantaneous SNR of $SS_{i,j}$
φ_i	the worst channel condition among all group members in MGroup i
$E_{i,j}^1$	The received signal power for $SS_{i,j}$ in Phase I
$E_{i,j}^2$	The received signal power for $SS_{i,j}$ from BS

$\bar{E}_{i,jB}$	The average received signal power of $SS_{i,j}$ from BS
$\bar{E}_{i,jk}$	The average received signal power of $SS_{i,j}$ from $SS_{i,k}$
N_0	The background noise power
R_i^1	The transmission rate of the BS in Phase I for MGroup i
R_i^2	The transmission rate of each cooperative transmitter in Phase II for MGroup i
C	Coverage ratio used for R_i^1
$Th_{i,j}^{CMS}$	The throughput of $SS_{i,j}$ for the proposed CMS scheme
$Th_{i,j}^{CON}$	The throughput of $SS_{i,j}$ for the conserve multicast scheduling scheme
Th_i^{CMS}	The group throughput of the MGroup i for the proposed CMS scheme
Th_i^{CON}	The group throughput of the MGroup i for the conserve multicast scheduling scheme