Dynamic Resource Allocation in Buffer-Aided Relay-Assisted Cellular Networks

by

Javad Hajipour

B.Sc., Electrical Engineering, Iran University of Science and Technology, Iran, 2005
M.Sc., Electrical Engineering, Sharif University of Technology, Iran, 2007

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF
THE REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

in

The Faculty of Graduate and Postdoctoral Studies

(Electrical and Computer Engineering)

THE UNIVERSITY OF BRITISH COLUMBIA
(Vancouver)

August 2015

Abstract

The increasing interest in wireless connectivity to the Internet has led to new technologies in cellular networks to provide ubiquitous access to users. One of the promising solutions is to deploy wireless relays, equipped with buffers, in different parts of the cellular networks, to improve both coverage and capacity. In this thesis, the goal is to investigate resource allocation in such networks, considering different challenges, system constraints and users' service requirements.

First, based on simple reasoning, analytical investigations and intuitive generalizations, we show that the use of buffers at relays improves both throughput and average end-to-end packet delay. Extensive computer simulations confirm the validity of the presented discussions and the derived results.

Subsequently, we propose Channel-, Queue-, and Delay-Aware (CQDA) resource allocation policies, which provide quality of service (QoS) for both delay-sensitive and delay-tolerant users, in a multiuser orthogonal frequency division multiple access (OFDMA) network enhanced with buffering relays. Numerical results demonstrate significant improvements in providing QoS through the proposed resource allocation policies compared with the existing algorithms.

Moreover, we introduce a perspective based on which we divide the network area, in a relay-assisted OFDMA system, to smaller cells served by the base station (BS) and the relays. Using convex optimization and dual decomposition, we derive closed form expressions for iterative signaling among the serving nodes to decide about resource allocation.

The resulted framework provides insights for designing efficient algorithms for practical systems.

Next, we introduce important parameters to be considered in the instantaneous problem formulation for data admission and resource allocation in the relay-assisted OFDMA cellular networks. Taking into account several practical constraints, we propose novel and efficient algorithms for deciding about time slot, subchannel and power allocation in a distributed manner. Numerical results confirm the effectiveness of the proposed parameters and algorithms in reaching the objectives and satisfying the constraints.

Finally, we propose a novel scheduling policy, which provides queue stability and is efficient and fair in terms of delay. The proposed policy can be used in the scenarios with shared or independent channels at the BS and relays, leads to less overhead, and facilitates decentralized resource allocation.

Preface

This thesis is based on the research I have conducted under the supervision of Dr. Victor Leung. The result of this research was several articles that have been either accepted or published, or are under review. I developed the ideas for these articles and wrote them under the supervision of Dr. Leung, who also helped in reviewing them. Except for the conference publication of Chapter 5, all the articles are co-authored also by Dr. Amr Mohamed from Qatar University. He provided valuable comments and feedback for the works and reviewed them for submission. For the article related to Chapter 2, Rukhsana Ruby helped in analysis as well as conducting the simulations and writing the manuscript. For the conference publication of Chapter 5, Dr. Ghasem Naddafzadeh helped in conducting the simulations and Dr. Peyman Talebifard helped in preparing the Introduction section of the article. In the following, the list of these publications are provided.

Publication related to Chapter 2

- Javad Hajipour, Rukhsana Ruby, Amr Mohamed and Victor C. M. Leung, "Buffer-Aided Relaying Improves Both Throughput and End-to-End Delay", *Submitted to a peer reviewed journal.*

Publication related to Chapter 3

- Javad Hajipour, Amr Mohamed and Victor C. M. Leung, "Channel-, Queue-, and Delay-Aware Resource Allocation in Buffer-Aided Relay-Enhanced OFDMA Net-

works," *Accepted for Publication in Trans. on Veh. Technol.*, Mar. 2015.

Publication related to Chapter 4

- Javad Hajipour, Amr Mohamed and Victor C. M. Leung, "Dynamic Distributed Resource Allocation in Relay-Assisted OFDMA Networks," in *Proc. of ICWMC*, June 2012.

Publication related to Chapter 5

- Javad Hajipour, Amr Mohamed and Victor C. M. Leung, "Utility-Based Efficient Dynamic Distributed Resource Allocation in Buffer-Aided Relay-Assisted OFDMA Networks," *Submitted to a peer reviewed journal.*

- Javad Hajipour, Ghasem Naddafzadeh, Peyman Talebifard and Victor C. M. Leung, "Power Efficient High-Rate Service Provisioning in Vehicular Networks," in *Proc. of ACM DivaNet*, Sep. 2014.

Publication related to Chapter 6

- Javad Hajipour, Amr Mohamed and Victor C. M. Leung, "Efficient and Fair Throughput-Optimal Scheduling in Buffer-Aided Relay-Based Cellular Networks," *Accepted for Publication in IEEE Commun. Lett.*, May 2015.

Table of Contents

Abstract ii

Preface iv

Table of Contents vi

List of Figures x

List of Abbreviations xiv

Acknowledgments xvii

Dedication xviii

1 Introduction 1

1.1 Related Works and Motivation 2

1.2 Research Questions and Objectives 6

1.3 Contributions 9

1.4 Thesis Organization 12

2 Buffer-Aided Relaying Improves Both Throughput and End-to-End Delay 14

2.1 Introduction 14

2.2 Background 15

2.3 Effect of Buffer Aided Relaying On the End-to-End Delay 17
2.3.1 Relaying Systems with Deterministic Data Arrivals and Bernoulli Channel Conditions 17
2.3.2 Relaying Systems with Bernoulli Data Arrivals and Channel Conditions 20
2.3.3 General Relaying Systems 25
2.4 Numerical Results 28
2.4.1 Bernoulli Data Arrivals and Channel Conditions 28
2.4.2 General Scenario 31
2.5 Summary 37

3 Channel-, Queue-, and Delay-Aware Resource Allocation 38
3.1 Introduction 38
3.2 System Model 40
3.3 Quality-Of-Service-Aware Resource Allocation Problem 44
3.3.1 The Main Objectives 44
3.3.2 Challenges of IRAP Formulation 45
3.4 Quality-Of-Service-Aware Cross Layer Scheduling and Resource Allocation 49
3.4.1 CQDA Policies 49
3.4.2 Enhanced CQDA Policies 59
3.4.3 Practical Considerations 61
3.5 Performance Evaluation and Discussion 62
3.6 Summary 78

4 Dynamic Distributed Resource Allocation Framework 79
4.1 Introduction 79
4.2 System Model 80

4.3 Cross Layer Scheduling and Resource Allocation 82
4.3.1 Problem Formulation 82
4.3.2 Dual Problem Formulation 85
4.3.3 Dynamic Distributed Resource Allocation 87
4.3.4 Solution of Main Dual Problem at the BS 88
4.4 Numerical Results 89
4.5 Summary 92

5 Utility-Based Efficient Dynamic Distributed Resource Allocation . . . 94
5.1 Introduction 94
5.2 Preliminaries 97
5.2.1 System Model 97
5.2.2 Stochastic Problem Formulation 100
5.2.3 Transformed Problem and Virtual Queues 102
5.3 Cross Layer Traffic Control and Resource Allocation 104
5.3.1 Instantaneous Problem 104
5.3.2 Traffic Control and Data Admission 106
5.3.3 Resource Allocation Challenges 107
5.3.4 Efficient Dynamic Distributed Resource Allocation 111
5.3.5 Efficient Dynamic Centralized Resource Allocation 119
5.4 Performance Evaluation and Discussion 120
5.5 Summary 130

6 Efficient and Fair Throughput-Optimal Scheduling 131
6.1 Introduction 131
6.2 Preliminaries 133
6.2.1 System Model 133

6.2.2 Background and Problem Statement 138
6.3 MMW Policy 141
6.3.1 Motivation and The Main Idea 141
6.3.2 Stability Analysis 145
6.4 Distributed and Semi-Distributed Implementations 146
6.4.1 Case1: Shared Channel 146
6.4.2 Case2: Independent Channels and Highly Scalable Framework . . . 147
6.5 Performance Evaluation 148
6.6 Summary 160

7 Conclusions and Future Work 162
7.1 Conclusions 162
7.2 Suggestions for Future Work 166

Bibliography 169

Appendices

A Assumptions for Channel Models 178

B Proof of Theorem 2.1 181

C Proof of Theorem 6.1 183

List of Figures

1.1 (a) Relay-assisted cellular network (b) OFDMA system 3

2.1 Simple queueing system 16

2.2 Queueing model for (a) conventional relaying system (b) buffer-aided relaying system; (c) joint channel conditions 19

2.3 Markov chain for the number of packets in the BS buffer 21

2.4 Average end-to-end packet delay in the case of Bernoulli channel distribution with $s_1 = s_2 = 0.9$ 29

2.5 Average end-to-end packet delay in the case of Bernoulli channel distribution with $s_1 = 0.5, s_2 = 0.9$ 30

2.6 Average end-to-end packet delay in the case of Bernoulli channel distribution with $s_1 = s_2 = 0.5$ 30

2.7 (a) BS queue size over time (b) relay queue size over time; at the arrival rate of 50 packets/slot. 32

2.8 CDF of end-to-end packet delays at the arrival rate of 50 packets/slot . . . 33

2.9 Effect of packet arrival rate at the BS on (a) average throughput in each time slot (b) average end-to-end packet delay. 34

2.10 (a) BS queue size over time (b) relay queue size over time; at the arrival rate of 100 packets/slot. 36

2.11 CDF of end-to-end packet delays at the arrival rate of 100 packets/slot . . 37

3.1 System model 41
3.2 Flowchart of the IRAP formulation 50
3.3 (a) Average PDR (b) CDF of the delay of the received packets; with 3 subchannels, 6 close and 6 far users, all delay-sensitive. 66
3.4 CDF of the delay of the received packets; with 3 allocatable subchannels from a pool of subchannels, 6 close and 6 far users, all delay-sensitive. . . . 68
3.5 (a) Average throughput (b) average queue size of the delay-tolerant users; with 4 subchannels, 6 close and 6 far users, half of the users in each group are delay-sensitive and the other half are delay-tolerant. 69
3.6 (a) Average PDR (b) CDF of the delay of the received packets for the delay-sensitive users; with 4 subchannels, 6 close and 6 far users, half of the users in each group are delay-sensitive and the other half are delay-tolerant. . . . 71
3.7 (a) Maximum packet delay of delay-sensitive users (b) average total throughput of delay-tolerant users; $N = 10$, $K = 22$, random distribution of users in the cell area. 73
3.8 Effect of data arrival rate of delay-sensitive users on (a) average PDR of delay-sensitive users (b) average total throughput of delay-tolerant users. . 75
3.9 Effect of data arrival rate of delay-tolerant users on (a) average PDR of delay-sensitive users (b) average total throughput of delay-tolerant users. . 77

4.1 System model 81
4.2 Similarity of the model to multicell network 84
4.3 CDF of system average throughput in each time slot, K=10 91
4.4 System average queue size over time, K=10 92
4.5 Effect of increase in the number of users on the system average throughput 93

5.1 System model 98

5.2 Effect of parameter I on (a) average power consumption of BS and relays (b) virtual power queue size of BS over time; $N = 7$, $|\mathcal{K}_0| = 6, |\mathcal{K}_m| = 1, m \in \mathcal{M}$. 122
5.3 CDF of (a) system utility (b) total overflow from the buffers of relays; at the data arrival rate of 20 packets/second 125
5.4 CDF of the system throughput at the data arrival rate of 20 packets/second 126
5.5 Effect of increase in the packet arrival rate at the BS on (a) the system average utility (b) average overflow from the buffers of relays 127
5.6 Effect of increase in the packet arrival rate at the BS on the system average throughput 128
5.7 Effect of increase in the packet arrival rate on the amount of data admitted for direct and indirect users 129
5.8 Effect of parameter W_e on providing fair data admission for direct and indirect users, at the arrival rate of 140 packets/second for every user 130

6.1 System model 134
6.2 (a) Parallel queues (b) combination of parallel and tandem queues. 142
6.3 Multihop relay-based cellular network. Square, circle and triangle represent the BS, RS and user, respectively. 149
6.4 (a) Percentage of time slots each queue in the RS exceeds the queue threshold (b) maximum queue sizes in the RS, over time, when $\hat{Q} = 14$ kbits. 151
6.5 (a) CDF of the average queue size for direct and indirect users (b) average bit delay of all the users; the case of shared channel. 153
6.6 CDF of Jain's fairness index for average bit delays of the users; the case of shared channel. 154

6.7 CDF of the average queue size for direct and indirect users; the case of shared channel, where direct users are located close to the BS and indirect users on the cell edge. 155

6.8 Average bit delay for direct and indirect users; the case of shared channel, where direct users are located close to the BS and indirect users on the cell edge. 156

6.9 (a) System average queue size over time (b) system average throughput in each time slot; the case of shared channel, where indirect users are located close to the RS. 157

6.10 (a) CDF of the average queue size for direct and indirect users (b) average bit delay of all the users; the case of independent channels. 159

6.11 CDF of Jain's fairness index for average bit delays of the users; the case of independent channels. 160

List of Abbreviations

AF	Amplify and Forward
BE	Best Effort
BER	Bit Error Rate
BS	Base Station
CDF	Cumulative Distribution Function
CF	Compress and Forward
CQDA	Channel-, Queue-, and Delay-Aware
CSI	Channel State Information
DDRA	Dynamic Distributed Resource Allocation
DF	Decode and Forward
EDCRA	Efficient Dynamic Centralized Resource Allocation
EDDRA	Efficient Dynamic Distributed Resource Allocation
EE	Energy Efficiency
EJUMR	Enhanced Joint Utility and Minimum Rate
ESUMR	Enhanced Separate Utility and Minimum Rate
FDRA	Frequency Domain Resource Allocation
FHDR	Fixed Half-Duplex Relaying
HoL	Head of Line
IRAP	Instantaneous Resource Allocation Problem

JUMR	Joint Utility and Minimum Rate
LTE	Long Term Evolution
LTE-A	LTE-Advanced
MDB	Maximum Differential Backlog
MSB	Maximum Sum Backlog
MMW	Modified Max-Weight
MW	Max-Weight
OFDMA	Orthogonal Frequency Division Multiple Access
PDR	Packet Drop Ratio
PF	Proportional Fair
QAM	Quadrature Amplitude Modulation
QCSI	Queue and Channel State Information
QoS	Quality of Service
QSI	Queue State Information
RC	Rate Constrained
RRM	Radio Resource Management
RS	Relay Station
RT	Real Time
SA	Subchannel Allocation
SE	Spectral Efficiency
SNR	Signal-to-Noise Ratio
SPAS	Subchannel and Power Allocation Strategy
SSD	Subchannel Sets Determination
STD	Slot Type Determination
SUMR	Separate Utility and Minimum Rate

TDS	Time Domain Scheduling
TPA	Total Power Adjustment
WiMAX	Worldwide Interoperability for Microwave Access

Acknowledgments

First of all, I would like to express my deep gratitude to my supervisor, Dr. Victor Leung, for providing the opportunity to study the PhD program at UBC, which helped me to learn a lot about my field of study and also about myself. I also want to thank him for his invaluable support and guidance over the past years, and for his patience on my delay to decide the research topic.

I would like to declare special thanks to Dr. Amr Mohamed for providing the opportunity to visit Qatar university for a period of time during my research. I am grateful for his comments and advices through my research which helped in preparing my thesis.

Also, I would like to extend my appreciation to my colleagues and friends for their helps and supports during the past years.

Most importantly, I feel indebted to my great family, my father, my mother and my brothers for their everlasting love and support not only during my studies at UBC but also throughout my entire life.

This work was supported by the Natural Sciences and Engineering Research Council (NSERC) of Canada.

Dedication

To my family

Chapter 1

Introduction

Cellular networks have experienced tremendous increase in the demand for wireless connectivity in the past decade. The exponential growth in the data introduced in the internet and the desire and need for accessing it, any where and any time, have shifted the wireless services from voice dominant to data dominant ones. In particular, the advent of smart phones has accelerated this trend due to their promising capabilities and growing market of mobile applications. In order to address this trend, industrial and standardization sectors have devised new solutions, through several generations of cellular networks such as 3G and 4G. In addition to allocating more spectrum, novel transmission techniques have been standardized considering the fact that the frequency resources are very scarce and expensive. Among those, Orthogonal Frequency Division Multiple Access (OFDMA) is the key enabling factor in evolution towards high-speed data services, which has been widely adopted in contemporary wireless cellular networks such as IEEE 802.16 Worldwide Interoperability for Microwave Access (WiMAX) and Long Term Evolution (LTE) [1–4]. In OFDMA, the system bandwidth is divided into multiple carriers (subchannels) and the data is transmitted through these parallel channels, leading to robustness against multipath fading, high spectral efficiency, multiuser diversity and flexibility in radio resource allocation. However, users who are far away from the Base Station (BS) or have blockages between the BS and themselves suffer from low data rates due to their weak wireless links. Wireless relaying is an attractive mechanism to overcome this limitation, which has gained significant attention among both industrial and academic researchers due to the

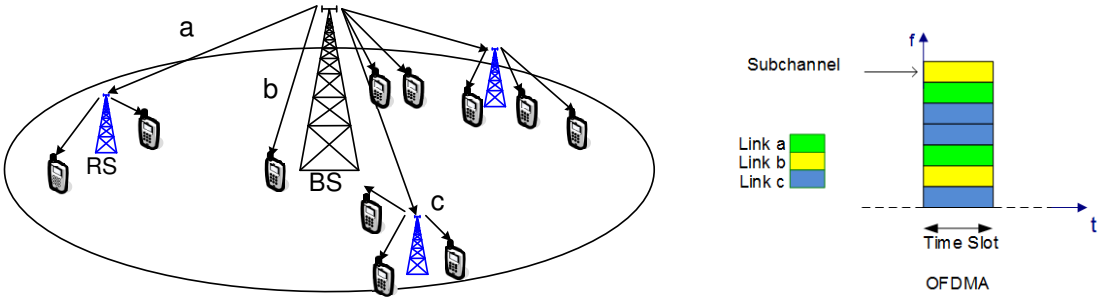
cost effective and fast deployment possibility of Relay Stations (RSs) [5, 6]. RSs use air interface for backhaul connectivity and are able to enhance the coverage, capacity, and energy efficiency without the need for costly wired backhaul deployment. They perform this through relaying protocols such as amplify-and-forward (AF), compress-and-forward (CF), or decode-and-forward (DF).

Scheduling and resource allocation are deciding factors for efficient use of wireless resources. Although the aforementioned advanced technologies lead to enhanced system capacity and coverage, intelligent scheduling and resource allocation algorithms are also needed to exploit the potentials of such systems and address the challenges in providing satisfactory service for users. In this thesis, our goal is to identify these opportunities and challenges and address them through suitable policies.

The rest of this chapter is organized as follows. In Section 1.1, we present the related work on resource allocation in relay-assisted wireless networks and discuss the motivation of this thesis. Section 1.2 discusses the research questions and objectives, and Section 1.3 briefly summarizes the contributions. Finally, Section 1.4 describes the thesis organization.

1.1 Related Works and Motivation

Resource allocation and scheduling are important issues in wireless networks due to the increasing demand of users for data traffic and the scarcity of radio resources. They become more complicated and challenging in relay-enhanced OFDMA networks, as the resources should be shared between the BS and relays [7, 8]. To clarify this, Fig. 1.1 shows a relay-assisted cellular network and an OFDMA system. The main resources in such systems are time slots, frequency subchannels and the power used on the subchannels. Depending on the objectives of the network operators and the constraints imposed by system limitations as well as the users' service requirements, different policies are needed for allocating these



(a) (b)

Figure 1.1: (a) Relay-assisted cellular network (b) OFDMA system

resources to different links of the network. There has been extensive research going on in this area in the past years and remarkable work has been done to address the challenges [9–31].

In particular, [9] suggested adaptive resource usage by utilizing access hop reuse and adaptive frame segmentation to improve the system capacity. In [10] cross layer scheduling in a relay network was studied as an optimization problem for maximizing the received goodput and, based on dual decomposition, a distributed algorithm was proposed. [11] proposed a distributed algorithm for power and subchannel allocation in a system with full-duplex and hybrid relaying methods. [12] studied the resource allocation problem in the presence of cognitive relays, and [13] proposed semi-distributed algorithm for resource allocation with low overhead and low computational complexity. In [14], the authors investigated the selection of the relays and subcarrier and power allocation for them, taking into account link asymmetry and imperfection in Channel State Information (CSI). Authors of [15] analyzed the effect of opportunistic scheduling and spectrum reuse on the system spectral efficiency and provided insights about the tradeoffs in the design of resource allocation and interference management algorithms. More other works studied frequency reuse schemes to improve the system capacity [16, 17]. In [18], authors considered subchannel reuse for transmissions from relays and based on a game-theoretic framework, they

designed distributed resource allocation schemes for transmissions from relays. The game theoretic approach is also used in [19–22] for channel allocation, interference control and joint consideration of relay selection and power allocation.

Providing quality of service (QoS) is another important concern for meeting the diverse requirements of user services in broadband wireless networks and has been addressed in [23–25]. [23] considered Best Effort (BE) and Rate-Constrained (RC) services and, based on convex optimization and dual problem formulation, introduced a QoS price concept for relay selection and subchannel allocation. In a similar service environment, authors of [25] studied joint optimization of the BS and relay power allocation in addition to relay selection and subchannel assignment; accordingly they introduced power and QoS prices and solved the problem using two level dual decomposition method. The work in [24], on the other hand, considered dynamic resource allocation for supporting the BE and Real-Time (RT) services simultaneously and, using a utility-based optimization problem, proposed an efficient relay selection and subchannel allocation algorithm to meet the stringent delay requirements of the RT users.

The common assumption in most of the literature in this area is that the relays do not have buffer to store packets for later transmission. Therefore, they have to operate in a "prompt" manner and forward their received data immediately in the following transmission interval. However, using relays that have buffering capability can make the resource allocation more flexible and enhance the system capacity. This has been investigated in several works and the throughput gain has been proved [32–38]. This gain is achieved due to the fact that the relay can store a user's data packets in a buffer when the channel condition between the relay and user is poor, and forward them when the channel becomes good. Motivated by this, several other works have considered the application of buffer-aided relaying in different areas [39–45].

Based on the above mentioned, the combination of OFDMA and buffer-aided relaying is a promising solution for ubiquitous high-data-rate coverage in cellular networks. However, any improvement in the system performance always comes at a cost. In the case of buffering relays, this cost has been investigated in [32–35, 38] and described as the increased delay of data packets in the relays' buffers. The arguments in these works are based on the assumption that the users have infinitely backlogged buffers in the BS, meaning that they always have data to transmit. However, this approach does not take into account the fact that in realistic scenarios, especially with the emerging data services, data packets arrive in a random and burst pattern at the BS buffers. Therefore, the effect of this on the packet delays needs more investigation to clarify the overall tradeoffs that need to be considered in the system design.

On the other hand, employing buffer-aided relays brings new challenges for scheduling and resource allocation in cellular networks and necessitates more investigations, to identify the issues and design sophisticated algorithms for addressing them. Recently there has been growing interest in this area [46–53]. [46] proposed a throughput-optimal policy, called Maximum Sum Backlog (MSB), to provide delay fairness in buffer-aided relay-assisted cellular networks. The authors of [47] considered quasi full-duplex relaying where a relay can receive and transmit simultaneously on orthogonal channels. Based the Queue and Channel State Information (QCSI) and using Max-Weight (MW) scheduling policy [54, 55], they proposed a fairness-aware resource allocation algorithm to provide ubiquitous coverage as well as load balancing. Similar work was done in [48] with the difference that it considered half-duplex relaying, and proposed iterative algorithms for transmissions over two consecutive subframes (BS transmission in the first one and relay transmission in the second one) using queue-length coupling. In [49] mobile-relay-enhanced networks were discussed and a channel-and-queue-aware algorithm was proposed to utilize the system

potentials. [51] studied joint time, subchannel and power allocation in LTE-A systems, where each time slot can either be used for transmissions on the BS-to-relays and BS-to-users links or the BS-to-users and relays-to-users links. [52] proposed a three step algorithm for joint optimization of long-term fairness and overall network throughput. In [53], authors considered a three node network, with one source, one relay and one destination, and based on Markov Decision Process (MDP), they studied the optimal link selection policy for maximizing the throughput in a system with finite relay buffer.

However, none of the above works considered service provisioning in buffer-aided relay-assisted wireless networks in the presence of users with stringent QoS requirements such as delay guarantees. Several other objectives and issues which already have been addressed in the relay networks without buffering, described earlier in this section, remain unanswered in the scenarios with buffering relays. In this thesis, we aim at identifying these issues and addressing them through sophisticated algorithms, which take into account the stochastic nature of data arrivals at the BS and RSs' buffers in addition to the randomness of wireless channel conditions. In the next section, we describe in detail the research questions and objectives addressed in this thesis.

1.2 Research Questions and Objectives

Research on resource allocation in buffer-aided relay-assisted cellular networks is new and needs time to address the challenges that arise in this area. One of the important issues is to identify the main tradeoffs related to using buffers at relays. This is important as it affects the decision on using relays with or without buffers in the practical systems. In the works in this area such as [32, 35, 38], only the queueing delay in relays is investigated and it is deemed that the use of buffers improves the throughput at the cost of increased delay. However, the question is what is the effect on the end-to-end packet delay, i.e.,

the delay that packets experience since their arrival at the BS buffer until delivery to the destination? Even though the packets experience delay in the relays' buffers, the queueing at the BS buffer should also be taken into account, as it affects the delay perceived by the end users. What are the tradeoffs considering this whole picture? We address this question in Chapter 2 and provide a ground for discussing the performance of resource allocation algorithms in buffer-aided relay-assisted cellular networks which are investigated in the subsequent chapters.

The main concerns of resource allocation in cellular networks can be classified into two categories. One is service oriented, which aims at efficient use of system resources and providing QoS for the users with specific service requirements. The other one is implementation oriented and is concerned about the efficiency and low-complexity in implementing the resource allocation algorithms.

To address the aforementioned concerns, we first consider service provisioning in buffer-aided relay-assisted networks for the users with stringent service requirements. In this context, we take the user satisfaction more important than the low-complexity concerns for the algorithms. We note that the resource allocation in a multiuser system for providing QoS is a complicated and challenging task, taking into account the strict delay requirements of delay-sensitive services and average throughput requirements of delay-tolerant ones. This is especially important due to the fact that the queuing delay in the relays needs to be taken into account in meeting the deadlines of the delay-sensitive packets. Because of this, the QoS-aware algorithms already proposed for prompt relays in the works such as [23–25] cannot be easily extended to the cases with buffering relays. Also, in the literature on resource allocation, the Instantaneous Resource Allocation Problem (IRAP) is usually assumed given and the goal is to design suitable algorithms for solving that. However, in a multihop network with buffering relays, the IRAP formulation for QoS provisioning is itself

a challenge. For this purpose, the definition of utility function, for average throughput provisioning, and the constraints imposed, for meeting packet transmission deadlines, need to be investigated in addition to the algorithm used later for solving the formulated problem. In Chapter 3 we discuss this in detail and propose sophisticated policies to address it.

Next, we consider the services with less stringent requirements and address the low-complexity and efficiency concerns about resource allocation algorithms. Noting the computational burden of these algorithms, it is always desirable to divide resource allocation tasks among the entities involved in serving the users, whenever service requirements of the users allow. A distributed resource allocation needs suitable framework for identifying the affecting variables and the required messaging between the serving entities. In the literature, there has been a lot of works for the scenarios with prompt relays. However, for buffer-aided relaying systems, it is still needed to derive such a framework taking into account the flexibilities brought by the use of buffers at the relays. We address this in Chapter 4.

Moreover, it is needed to take into account the constraints that might be imposed on resource allocation in practical systems. In particular, the time limitations for deciding about the allocation of the resources in the presence of fast varying channel conditions should be taken into account in designing efficient distributed algorithms. While the distributed frameworks derived based on mathematical tools provide an insight on the affecting factors, low-complexity resource allocation algorithms are necessary for practical considerations when the users' services are not very challenging as in the case of BE services. Furthermore, fair data admission for BE services and also constraints on average power consumption of the serving nodes in cellular networks require careful attention to the resource allocation problem formulation. Even though the Lyapunov drift-plus-penalty policy, studied in [51, 55], provides a useful framework for data admission and satisfying

the constraints defined in average sense, it needs more consideration for use in cellular networks. We investigate these issues and challenges in Chapter 5.

Finally, noting that the users connected to relays experience more delay than the ones connected to the BS, it is important to design efficient scheduling algorithms for providing fairness in terms of delay among the users with different number of hops. The well known MW scheduling policy [54, 55], which is usually used for stabilizing the queues and achieving the optimum throughput, has been mostly considered in multihop mesh networks where a packet can be routed through different paths to reach the destination. However, in the relay-based cellular networks with one path for packet transmission from the source BS to the destination, MW can lead to discrimination among the users with one hop and two hops distance from the backbone network, in terms of delay. Although MSB [46] tries to address this, it can lead to instability in the scenarios with shared channels for the BS and relays. Therefore, new policies are needed to address fairness in terms of delay in multihop cellular networks. We study this in Chapter 6.

1.3 Contributions

In this section, we outline the contributions in this thesis to address the above mentioned issues. The main contributions are summarized as follows:

- In Chapter 2, we provide insights on the effect of buffering relays on the end-to-end delay of users' data, from the time they arrive at the source until delivery to the destination. We also analyze end-to-end packet delay in the relay networks with Bernoulli data arrivals and channel conditions, and prove that the data packets experience lower average end-to-end delay in buffer-aided relaying system compared with the conventional one. Furthermore, using intuitive generalizations, we clarify that the use of buffers in relays improves not only throughput, but ironically the average

end-to-end packet delay. Through computer simulations, we validate our analytical results for the systems when the data arrival and channel condition processes follow Bernoulli distribution. Moreover, via the extensive simulations under the settings of practical systems, we confirm our intuition for general scenarios.

- In Chapter 3, we propose novel Channel-, Queue-, and Delay-Aware (CQDA) policies for providing QoS in OFDMA networks enhanced with buffering relays. CQDA policies take into account the QoS requirements of both delay-sensitive users with the goal of meeting packet deadline constraints, and delay-tolerant users who need guarantees on their average throughput. We provide a framework for "time domain scheduling" and "frequency domain resource allocation", based on which, the proposed CQDA policies formulate the IRAP and decide about routing and resource allocation. These policies take different approaches to determine the set of users considered in the utility function, the delay budget division between the BS and relays, the routing path of delay-sensitive users' packets as well as the values of minimum rate requirements for serving their queues. At the end, they use an iterative algorithm to solve the resulted IRAPs. Numerical results show significant improvements in throughput and delay performance of the proposed resource allocation mechanisms compared with the existing algorithms.

- In Chapter 4, we propose a novel framework for distributed resource allocation in a buffer-aided relay-assisted OFDMA system. This framework models the network as a multicell scenario with small serving areas where each of the relays and BS serves one of these areas using shared subchannel and power resources. It provides an insight for reducing signaling overhead and computation burden on the BS in practical systems with buffering relays. We formulate the power and subchannel allocation problem as a convex optimization problem and propose Dynamic Distributed Resource Allocation

(DDRA) algorithm, where the BS and relays decide about the allocation of the system resources by passing messages among themselves and based on the QCSI. Simulation results confirm that the proposed algorithm is able to utilize the system resources well, make the system queues stable and lead to high throughput.

- In Chapter 5, we propose effective parameters for instantaneous problem formulation, which adapt the well-known Lyapunov drift-plus-penalty policy to buffer-aided relay-assisted cellular networks. One of these parameters is the extra weight for the links of the relayed users from the BS and relays. The other parameter is the importance coefficient for the virtual power queues corresponding to the constraints on the average power consumption of the BS and relays. These parameters enable fair data admission for BE services and facilitate satisfying the average power constraints. Moreover, we identify the challenges that arise even when equal power allocation of the BS and relays is considered on the subchannels of an OFDMA system. We introduce a low-complexity strategy to break the ties in power and subchannel allocation. Using that, we design efficient and low-complexity distributed and centralized resource allocation methods for buffer-aided relay-assisted OFDMA networks, which take into account several practical constraints. Specifically, the proposed Efficient Dynamic Distributed Resource Allocation (EDDRA) scheme is suitable for use in practice as it imposes less overhead on the system and splits the resource allocation tasks among the BS and relays. Extensive simulation results show the effectiveness of the proposed parameters in meeting the objective and the constraints of the studied problem. We also show that the proposed EDDRA scheme has close performance to the proposed centralized one and outperforms an existing centralized algorithm.

- Finally, in Chapter 6, we propose novel throughput-optimal scheduling policy which stabilizes the system queues and at the same time is efficient and fair in terms of

queueing delay. We show that MW or MSB, proposed in the literature, are either unfair or unstable in the shared channel scenarios. We modify MW policy and propose a new version of throughput-optimal algorithms which we refer to as Modified Max-Weight (MMW). In MMW, by defining a suitably large threshold, a link's weight is proportional to just the corresponding local queue size either in the BS or RS, almost all the time. This makes MMW suitable for use in both shared and independent channel scenarios, with either centralized or decentralized network implementations. MMW alleviates the need to report any information about the queue sizes of relayed users, almost all the time. Also, it can adjust a parameter to further improve delay fairness between the relayed and direct users. Numerical results confirm that MMW leads to similar delay for direct and relayed users and also has low signaling overhead.

1.4 Thesis Organization

This thesis is organized as follows. First we discuss the effect of buffering relays on the end-to-end packet delay, through mathematical analysis and intuitive generalizations in Chapter 2. The goal is to dispel the concern that the use of buffer in relays would increase the delay in the system. In Chapter 3, we discuss the challenge in IRAP formulation for providing QoS in the presence of users with heterogeneous service requirements. We propose novel CQDA policies for deciding about the parameters of utility function and the problem constraints. In Chapter 4, we provide a novel perspective for resource allocation and use convex optimization to derive closed form equations for distributed power and subchannel allocation. Data admission control and efficient distributed resource allocation is proposed in Chapter 5, where we take into account several practical constraints. Chapter 6 discusses efficient throughput-optimal algorithms in relay-based cellular networks, taking into account the fairness in terms of queueing delay. The conclusion and some potential

future work are presented in Chapter 7. Finally, the Appendices present the assumptions for the channel models and the proofs for the theorems.

Chapter 2

Buffer-Aided Relaying Improves Both Throughput and End-to-End Delay

2.1 Introduction

Wireless relays are promising solutions for enhancing the capacity and coverage of cellular networks. Usually in the literature in this area, it is assumed that the relaying is performed in two consecutive subslots of a transmission interval; i.e., in the first subslot, the BS transmits to the relay and in the second one, the relay forwards the received data to the destination. Recently it has been shown that using the buffering technique at the relay can improve the system throughput [32, 35, 38, 56]. This is achieved due to the fact that the buffering capability allows the relay to store the packets when the channel condition is bad and transmit when it is good. The drawback for this capability is usually deemed to be the increase in the packet delays due to queueing in the relay, and the works in [32, 35, 38, 56] have tried to investigate and discuss the trade off between throughput and delay. These investigations are based on the assumption of infinitely backlogged buffers in the source, i.e., the BS, and consider the queueing delay only at the relay buffer without taking into account the queue dynamics at the BS. However, the perceived delay at the destination is

affected by the queueing both at the BS and the relay. Therefore, it is needed to investigate the delay since the packets arrive at the BS until delivery to the destination.

In this chapter, we aim at filling the above mentioned gap and clarifying the tradeoffs in using buffer-aided relaying. For this, we first present simple reasoning and discuss the cause of queue formation in a simple queueing system. Based on that we provide an insight on the delay performance in buffer-aided and conventional relaying. Then, we study the delay performance in the case of Bernoulli data arrivals and channel conditions and derive the closed form equations for average end-to-end packet delay, based on which we prove that the average end-to-end packet delay in buffer-aided relaying is in fact lower than that in the conventional relaying. Then, we discuss general scenarios and through intuitive generalizations, we conclude that the buffering relays in fact improve throughput as well as the average end-to-end packet delay. Using simulations, we verify our analysis and demonstrate the validity of the presented perspective. To the best of our knowledge, this is the first work that discusses the effect of buffering relays on the overall waiting time in a relaying network and provides the aforementioned insight and conclusion.

The rest of this chapter is organized as follows. Section 2.2 provides a background on the queueing delay based on a simple queueing system. In Section 2.3 we study end-to-end delay performance of conventional and buffer-aided relaying networks. Section 2.4 provides numerical results and finally, the conclusion is presented in Section 2.5.

2.2 Background

In this section, we study a simple queueing system and discuss the cause of packet delays, to provide a basis for the next section which studies the end-to-end packet delay in relaying networks.

Let consider a single buffer, as shown in Fig. 2.1, which is fed by a deterministic data

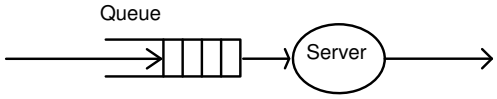


Figure 2.1: Simple queueing system

arrival process and served by a single server. We assume that time is divided into slots with equal lengths, indexed by $t \in \{1, 2, ...\}$. The total number of data packets that arrive at the buffer is N. Starting from $t = 1$, one packet arrives per time slot. Therefore, the last packet arrives at $t = N$. For simplicity, we assume that the arrivals occur at the beginning of time slots. The server might be active or inactive in each time slot. When it is active, it can serve only one packet per time slot, where the service implies delivering the packet to the destination.

We note that if the server is active in each time slot $t \in \{1, ..., N\}$, each packet will be served immediately after its arrival. In this case, there is no queue formed in the buffer and consequently, each packet experiences an overall delay of one time slot, which is due to the time spent in the server. Accordingly, the packets will arrive at the destination at the beginning of time slots $t \in \{2, ..., N + 1\}$. However, if the server is inactive in the first time slot, the first packet has to wait in the buffer until time slot 2, to get served. Then, in time slot 2, when the second packet arrives, the server is busy with serving the first packet. Therefore, the second packet also experiences one slot delay in the queue and one slot delay in the server. In the similar manner, all the following packets incur the same queueing and service delays. In other words, the delayed operation of the server causes the nonzero queueing delay for the first packet, which is transferred to the subsequent packets as well.

Based on the above discussion, if the server is inactive in time slot $x \in \{1, ..., N\}$, it adds one slot to the queueing delay (and the overall waiting time) of every packet arrived in slot x or afterward. In general, the packet which arrived in time slot t will experience

a queueing delay of n_t and will be delivered in time slot $t + n_t + 1$, where n_t indicates the number of slots before and including t in which the server was inactive. It is clear that the cause of queue formation in such systems is the interruption in the operation of the server, which is translated to queueing delays of the data packets.

2.3 Effect of Buffer Aided Relaying On the End-to-End Delay

In this section, first we consider that data arrives in a deterministic manner and the availability of the channels follows Bernoulli distribution, and provide an insight on the end-to-end delay performance for conventional and buffer-aided relaying systems. Then, we analytically derive the end-to-end delay for these systems, where both the data arrival process and the availability of the channels follow Bernoulli distribution. Finally, we discuss general cases and present the intuitions about the end-to-end delay performance. Note that in all the following discussions it is assumed that the channel conditions of the BS and relay are uncorrelated.

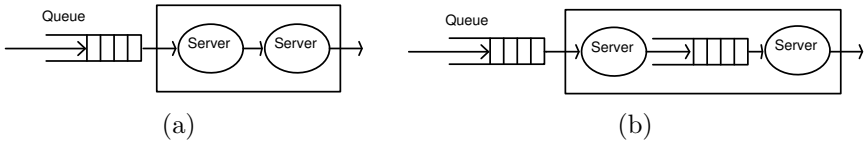
2.3.1 Relaying Systems with Deterministic Data Arrivals and Bernoulli Channel Conditions

Let consider a relay network, with one source node, i.e., the BS, one relay node and one destination (or user) node, where the relay works based on DF technique. It is assumed that there is no direct link between the BS and the user, and the transmissions are done only through the relay. There is only one channel in the system, which can be used for either transmissions from the BS to the relay or from the relay to the user. Each time slot is divided into two subslots, where the BS and relay can transmit in the first and

second subslots, respectively. We use c_1 and c_2 to indicate the BS channel condition (for the link between the BS and relay) and relay channel condition (for the link between the relay and user), respectively. These variables can either be "Good" or "Bad" , meaning respectively that it is possible to transmit one or zero packet on the corresponding channel. The probability of being "Good" is s_1 and s_2 for the BS and relay channel, respectively. Fig. 2.2(a) shows the queueing model for a conventional relaying system, where the relay does not have buffer, and therefore, it has to transmit its received data immediately in the next subslot. The server 1 and server 2 indicate the wireless channel from the BS to relay and from the relay to user, respectively. On the other hand, Fig. 2.2(b) indicates a relaying network, where the relay has a buffer which allows it to store the data packets and transmit whenever its channel is good. In both of the figures, the rectangle enclosed around the servers is to abstract the overall serving behavior of the system from the time that the BS starts to transmit data packets until their delivery to the user. Note that the works in [32, 35, 38, 56], in fact study the delay by considering only the time a packet spends inside this rectangle, and do not take into account the waiting time in the BS queue, which occurs before the transmission from the BS to the relay.

In the following, we consider the data arrivals in the BS buffer as the deterministic process, with N packets, mentioned in the previous subsection. Taking the overall service behavior of the systems into account, we discuss the overall waiting time of data packets in both the conventional and buffer-aided relaying systems. The overall waiting time is in fact the end-to-end delay, from the time that a packet arrives at the BS buffer until it is delivered to the user.

Fig. 2.2(c) shows the different states for the joint conditions of the BS and relay channels, in which G and B indicate "Good" and "Bad" conditions, respectively. We note that the system with conventional relaying serves the packets only when $c_1c_2 = GG$, and with



(a) (b)

c_1c_2	probability
GG	s_1s_2
GB	$s_1(1-s_2)$
BG	$(1-s_1)s_2$
BB	$(1-s_1)(1-s_2)$

(c)

Figure 2.2: Queueing model for (a) conventional relaying system (b) buffer-aided relaying system; (c) joint channel conditions

the probability of $s = s_1s_2$. In the other three cases, i.e., when either or both of c_1 and c_2 are "Bad", the packets remain in the BS buffer and are not transmitted. Therefore, based on the discussions in the previous subsection, the overall server in the system is inactive with the probability of

$$u_{nb} = P(GB) + P(BG) + P(BB) = 1 - s = 1 - s_1s_2 \tag{2.1}$$

where u_{nb} indicates the interruption probability for the overall server in the system without buffering in the relay. Considering this, in each time slot, the probability of "increase of one slot" in the overall waiting time of the packets present in that time slot or arrived after that is $u_{nb} = 1 - s_1s_2$. Here, the increase in the overall waiting time is due to the increase in the BS queueing delay of those packets.

Now consider the system where the relay has a buffer. We note that if the channel conditions are as BB in time slot x, similar to the system with conventional relaying, there will be an increase of one slot in the overall waiting time of the packets present in the time slot x or arriving afterward. However, for the channel conditions as GB and BG, the case is different. In order to clearly investigate these states, first we consider the following example:

- In time slot $t = 1$, the channel conditions are as GB. Therefore, in the first subslot,

packet 1 will be transmitted from the BS to relay; but due to the "Bad" channel condition of relay, it will not be transmitted to the user in the second subslot and will be stored in the relay's buffer.

- In time slot $t = 2$, the channel conditions are as BG. Therefore, in the first subslot, there will not be any transmission from the BS to relay and the overall waiting time of the packets $2, ..., N$ will be increased by *one* slot. However, due to good condition of the relay channel, packet 1 will be transmitted from the buffer of the relay to the user in the second subslot.

In the above example, it is observed that packet 1 is served by the relay in time slot $t = 2$ and therefore, it is delivered to the user at time slot $t = 3$. This has become possible due to the queueing of that packet in the relay's buffer. Note that with conventional relaying, however, in the above example, packet 1 would remain in the BS queue in both time slots $t = 1$ and $t = 2$, and the overall waiting time would increase by *two* slots for *all* the packets. Based on the above discussion and considering the nonzero probability of having channel conditions as GB and BG in two consecutive time slots, it can be concluded that $u_b < u_{nb}$, where u_b is the interruption probability of the overall server in the buffer-aided relaying system. In other words, the buffering capability in the relay reduces the overall waiting time for the data packets. This is achieved due to the fact that the queue size in the BS is reduced, and the data packets transferred to the relay buffer enable the efficient use of the relay channel.

2.3.2 Relaying Systems with Bernoulli Data Arrivals and Channel Conditions

Now, we consider relaying networks where both data arrivals and channel conditions follow Bernoulli distribution. We assume that in each time slot, the probability of one packet

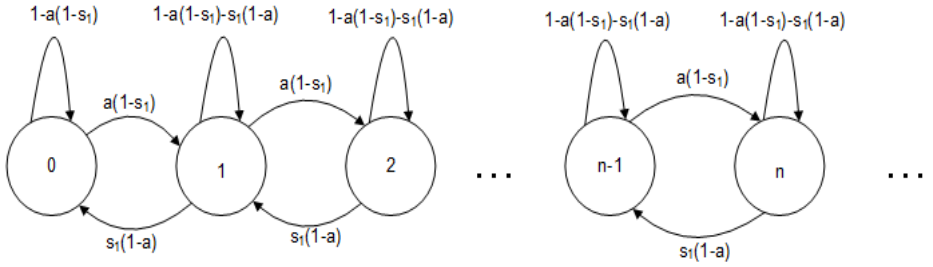


Figure 2.3: Markov chain for the number of packets in the BS buffer

arrival at the BS buffer is a, and, as before, the probability of "Good" channel condition for the BS and relay is equal to s_1 and s_2, respectively. It is assumed that $a < s_1 s_2$ and therefore, the system queues are stable in the case of conventional and buffer-aided relaying [55, Chapter 2]. In the following, when we use subscript b and nb for the variables, we refer to them in the case with buffering and without buffering in the relay, respectively.

Buffer Aided Relaying System

Based on [57, Section 7.5], Fig. 2.3 shows the Markov chain model for the queue dynamics at the BS buffer for the buffer-aided relaying network, where each state represents the number of packets in the queue. Let p_n, $n \in \{0, 1, \cdots\}$, denote the probability that in steady state, there are n packets in the BS queue. Note that due to equilibrium in the steady state, we have:

$$\begin{aligned} p_0 &= [1 - a(1 - s_1)]\, p_0 + s_1(1 - a)p_1, \\ p_n &= a(1 - s_1)p_{n-1} + [1 - \{a(1 - s_1) + s_1(1 - a)\}]\, p_n + s_1(1 - a)p_{n+1}, n = 1, 2, \cdots \end{aligned}$$

Based on the above equations, the probability of each state can be written as

$$p_n = \rho^n p_0, \;\; n = 0, 1, 2, \cdots, \tag{2.2}$$

where

$$\rho = \frac{a(1-s_1)}{s_1(1-a)}. \tag{2.3}$$

Considering the fact that $\sum_{n=0}^{\infty} p_n = 1$, we have:

$$p_0 = 1 - \rho. \tag{2.4}$$

Therefore, when a new packet arrives at the BS buffer, the expected number of packets it will see in the queue can be expressed by

$$E(Q_b^B) = \sum_{n=0}^{\infty} n p_n = \frac{\rho}{1-\rho} = \frac{a(1-s_1)}{s_1 - a}. \tag{2.5}$$

Note that when a new packet arrives at the BS buffer, its expected delay until departing can be split into two parts. The first part is the expected time that it has to wait until the packets already in the queue are served, i.e., $E(Q_b^B)E(T_b^B)$, where $E(T_b^B)$ is the expected delay imposed due the service of each packet when it is in the head of queue. The second part is the expected time since the packet itself gets to the head of the queue until its service is completed, which is denoted as $E(T_b^{B*})$. Therefore, the waiting time of a packet in the BS, $E(D_b^B)$, can be written as $E(D_b^B) = E(Q_b^B)E(T_b^B) + E(T_b^{B*})$. This is in fact the well known *mean value approach* which holds for queueing systems with memoryless data arrival processes [58, Section 4.3].

The interpretation of $E(T_b^B)$ is as follows. The delay caused due the service of a packet in the head of the queue is 1 slot with the probability of s_1 (this is in the case that the BS channel is good at the time that the packet gets to the head of queue). It is $(1+1)$ slots with the probability of $(1-s_1)s_1$, $(2+1)$ slots with the probability of $(1-s_1)^2 s_1$, $(k+1)$ slots with the probability of $(1-s_1)^k s_1$, and so on. Therefore, the expected delay caused

due the service of a packet in the head of queue is given by

$$\begin{aligned} E(T_b^B) &= s_1 + (1+1)(1-s_1)s_1 + (2+1)(1-s_1)^2 s_1 + \cdots \\ &= \sum_{k=0}^{\infty} (1-s_1)^k s_1 .(k+1) \\ &= s_1 \sum_{k=0}^{\infty} (1-s_1)^k k + s_1 \sum_{k=0}^{\infty} (1-s_1)^k \\ &= s_1(1-s_1)\left[\frac{d}{ds_1}\left(-\sum_{k=0}^{\infty}(1-s_1)^k\right)\right] + s_1 \frac{1}{1-(1-s_1)} \\ &= -s_1(1-s_1)\frac{d}{ds_1}\frac{1}{s_1} + s_1\frac{1}{s_1} \\ &= \frac{1-s_1}{s_1} + 1 \\ &= \frac{1}{s_1} \end{aligned} \tag{2.6}$$

On the other hand, we can compute $E(T_b^{B*})$ as follows. Considering that the packet is at the head of the queue, its delay until the departure from the BS is equal to 0.5 with the probability of s_1, $(1+0.5)$ with the probability of $(1-s_1)s_1$, $(2+0.5)$ with the probability of $(1-s_1)^2 s_1$, $(k+0.5)$ with the probability of $(1-s_1)^k s_1$, and so on. Hence, the expected waiting time of the packet when it has no queue in the front is

$$\begin{aligned} E(T_b^{B*}) &= 0.5s_1 + (1+0.5)(1-s_1)s_1 + (2+0.5)(1-s_1)^2 s_1 + \cdots \\ &= \sum_{k=0}^{\infty} (1-s_1)^k s_1 .(k+0.5) \\ &= s_1 \sum_{k=0}^{\infty} (1-s_1)^k k + 0.5 s_1 \sum_{k=0}^{\infty} (1-s_1)^k \\ &= \frac{1-s_1}{s_1} + 0.5. \end{aligned} \tag{2.7}$$

Based on the above discussions, the expected delay of a packet in the BS is equal to

$$\begin{aligned} E(D_b^B) &= E(Q_b^B)E(T_b^B) + E(T_b^{B*}) \\ &= \tfrac{a(1-s_1)}{s_1-a}\tfrac{1}{s_1} + \tfrac{1-s_1}{s_1} + 0.5 \\ &= \tfrac{1}{s_1}\left[\tfrac{a(1-s_1)}{s_1-a} + 1\right] - 0.5 \\ &= \tfrac{1-a}{s_1-a} - 0.5. \end{aligned} \tag{2.8}$$

We note that in each time slot, either one or zero packet departs the BS. Therefore, the packet departures from the BS can be modeled as a Bernoulli process. Due to the stability of the queues, the data departure rate from the BS is equal to the data arrival rate in its buffer. Consequently, the probability that one packet departs the BS, or, equivalently, the probability that one packet arrives at the relay buffer is equal to a. As a result, the average delay that a packet experiences in the relay can be computed in the similar manner as the average delay in the BS buffer, which is expressed by

$$E(D_b^R) = \frac{1-a}{s_2-a} - 0.5. \tag{2.9}$$

Based on (2.8) and (2.9), the average waiting time of a packet in the buffer-aided relaying system is given by

$$E(D_b) = E(D_b^B) + E(D_b^R) = \frac{1-a}{s_1-a} + \frac{1-a}{s_2-a} - 1. \tag{2.10}$$

Conventional Relaying System

Note that in the conventional relaying system, there is no buffering delay at the relay and the service probability for serving the BS buffer is $s_1 s_2$. Therefore, the average number of packets in the BS can be obtained by replacing s_1 with $s_1 s_2$ in (2.5). Similarly, the average delay caused for a packet due to the service of the packets in front of it can be

computed based on (2.6) and by using $s_1 s_2$ instead of s_1. On the other hand, the delay that a packet experiences when it gets to the front of the queue can be obtained based on (2.7) and by replacing 0.5 with $0.5 + 0.5$. This is because it takes the whole time slot for the packet to be transmitted from the BS to the destination. Considering these, the end-to-end delay in the conventional relaying system can be expressed by

$$E(D_{nb}) = \frac{1-a}{s_1 s_2 - a}. \tag{2.11}$$

In order to compare the delay performance of the conventional and buffer-aided relaying systems, Theorem 2.1 states and proves the main result of this subsection.

Theorem 2.1 Consider a relaying network where the data arrival process at the BS and the channel availability process follow Bernoulli distribution. Then, the average end-to-end packet delay in the buffer-aided relaying system is less than or equal to that in the conventional one. In other words, we have:

$$E(D_b) \leq E(D_{nb}), \tag{2.12}$$

where the equality holds only in the case that the channels are always in "Good" condition, i.e., $s_1 = s_2 = 1$.

The proof of Theorem 2.1 is given in Appendix B.

2.3.3 General Relaying Systems

Now consider a general scenario, where the data arrival and channel condition processes follow general distributions. We use $r_{br}(t)$, $r_{rd}(t)$, and $r_{bd}(t)$ to show the achievable transmission rate in time slot t between the BS and relay, the relay and destination, and the

BS and destination, respectively. Without buffering, the BS needs to transmit to the relay in the first subslot and then, the relay has to forward it immediately in the next subslot. We know that in this case, the end-to-end achievable rate between the BS and the user is $r_{bd}(t) = \frac{1}{2}\min\{r_{br}(t), r_{rd}(t)\}$. Due to this, the transmission rate in each slot is limited by the link with the worst channel condition in that time slot.

However, when the relay has a buffer, there is no necessity for the immediate forwarding of the data and the above mentioned limitation is relaxed; therefore, the BS has the opportunity for transmitting continuously to the relay when the channel condition from the BS to relay is good. Then, the relay can store them in the buffer to transmit when the channel from the relay to user is good. Because of this, the buffering makes it possible to improve the system throughput [32, 35, 38, 56]. Improvement in the throughput is equivalent to the improvement in the end-to-end service rate of the data arrived at the BS buffer. In other words, the increase in the system throughput means that more data is transferred from the BS to the user, or equivalently, the same data is transferred from the BS to the user in a less amount of time. Therefore, on average, packets experience lower end-to-end delay, i.e., the delay since their arrival at the BS until delivery to the destination.

Based on the above discussion, we make the conclusion as follows. Although buffer-aided relaying results in queueing delay in the relay, it also facilitates data transfer from the BS to the user and leads to a large reduction in the queueing delay at the BS. Therefore, the overall effect is the improvement of the average end-to-end packet delay. In summary, we state this as follows.

Proposition: Using buffer at the relay improves the system throughput, and therefore, it reduces the average end-to-end packet delay.

We note that when buffering is used in the relay, a scheduling policy is required to stabi-

lize the system queues. Specifically, in each subslot, this policy should decide on allocating the channel to the BS or relay such that the system queues remain bounded. For this, throughput-optimal algorithms should be considered. A scheduling policy is throughput-optimal if it makes the system queues stable, if the stability is feasible at all with any other policy [59]. Note that this definition assumes infinite capacities for the system buffers and takes into account the fact that data arrive finitely and in a random pattern at the BS buffer. Therefore, having stable queues ensures that the average data departure rates of all the buffers are equal to their average data arrival rates and consequently, the arrived packets at the BS are delivered to their destination with a finite average delay [60]. As a result, the maximum possible throughput is obtained which is equal to the average data arrival rate at the BS (assuming no packets are lost or dropped). An important throughput-optimal policy in wireless networks with fixed number of queues and stationary ergodic data arrival processes is the well-known MW method [54, 55, 60]. MW aims at maximizing the weighted rates of the links, where the weight of a link is considered proportional to the difference in the queue sizes at the two ends of the link[1]. MW is an attractive scheduling policy for stabilizing the queues in buffer-aided relay networks as it works by utilizing just the instantaneous QCSI and does not require information about the probability distribution of packet arrival processes and channel states. Considering the above mentioned, we summarize the costs of buffer-aided relaying in the following remark.

Remark: Note that the costs for the improvements brought by buffer-aided relaying are the requirement for a memory to buffer data at the relay, and the necessity for a scheduling algorithm to keep the queues stable.

[1]It is assumed that the data packets exit the system in the destination and therefore, the queue size at the destination is zero.

2.4 Numerical Results

To verify the presented discussions, we have conducted extensive Matlab simulations over 10000 time slots. We have investigated the cases that the data arrival and channel condition processes follow Bernoulli distribution, as well as general cases with the settings of a practical system. We present the simulation results in the following and discuss the effect of buffer-aided relaying on the end-to-end packet delay. As stated before, the end-to-end delay of a packet is considered as the time elapsed since the packet arrives at the BS buffer until delivery to the user.

2.4.1 Bernoulli Data Arrivals and Channel Conditions

In order to validate the analysis provided in subsection 2.3.2, in Figs. 2.4, 2.5 and 2.6 we present the average packet delay obtained from both the mathematical analysis and the simulation. In each of these figures, we have fixed the values of s_1 and s_2 and have evaluated the effect of increase in a on the average end-to-end packet delay. In order to maintain the stability of the system queues, we have considered $a < s_1 s_2$. The graphs of mathematical analysis are plotted using (2.10) and (2.11). On the other hand, the graphs of simulation results are plotted based on the delays of packets in the simulated conventional and buffer-aided relaying systems with Bernoulli data arrivals and channel conditions. Fig. 2.4 displays the case of high probability for good channel conditions at the BS and relay, i.e., $s_1 = s_2 = 0.9$. It is clear that the simulation results are quite close to the analytical ones, which confirms the validity of the mathematical analysis. Moreover, the results confirm that the buffer-aided relaying has lower packet delays compared with the conventional relaying. As expected, both of the systems incur larger delay as the packet arrival probability increases. However, the delay in the conventional relaying increases faster compared with that in the buffer-aided relaying.

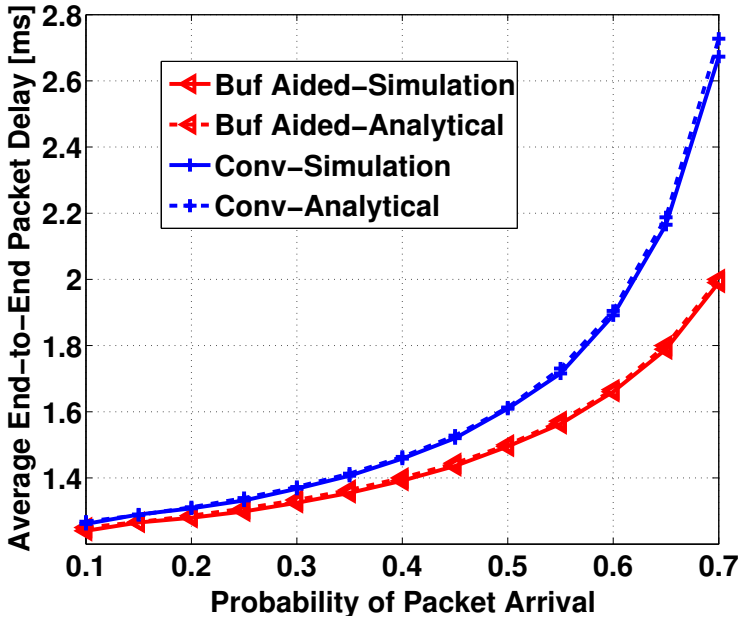


Figure 2.4: Average end-to-end packet delay in the case of Bernoulli channel distribution with $s_1 = s_2 = 0.9$

Furthermore, Fig. 2.5 and Fig. 2.6 show the results for the cases that either or both of the channels have relatively lower probability of being in good condition. It is observed that when the channels have lower probability of being in good condition, the conventional relaying results in significantly higher delays even at the lower data arrival rates. In particular, the performance difference of these relaying systems is larger in Fig. 2.5 compared with Fig. 2.4 and the largest in Fig. 2.6. This is because when the probability of good channel conditions is low, in the case of conventional relaying, the BS has to wait for a long time before having both the channels favorable for transmission. However, in the case of buffer-aided relaying, the BS can transmit to the relay even when the relay channel is bad. Then, the relay can buffer the received data and transmit in its subslots whenever its channel is good.

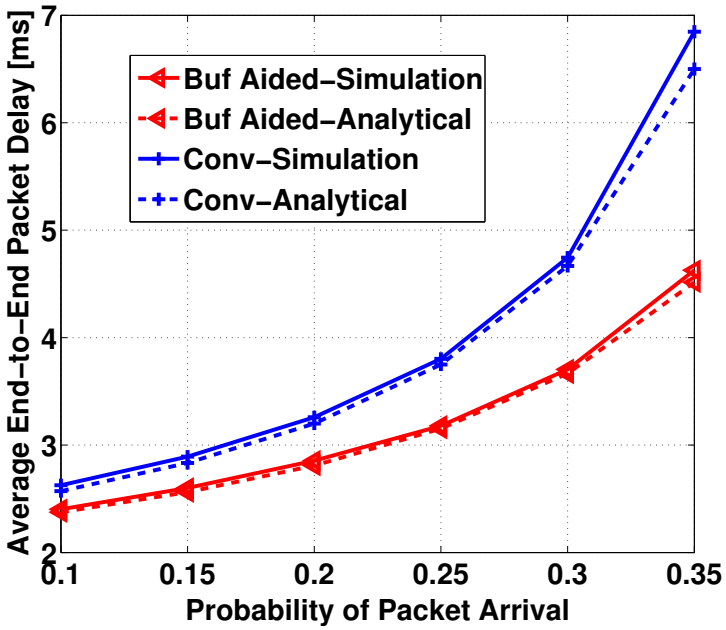


Figure 2.5: Average end-to-end packet delay in the case of Bernoulli channel distribution with $s_1 = 0.5, s_2 = 0.9$

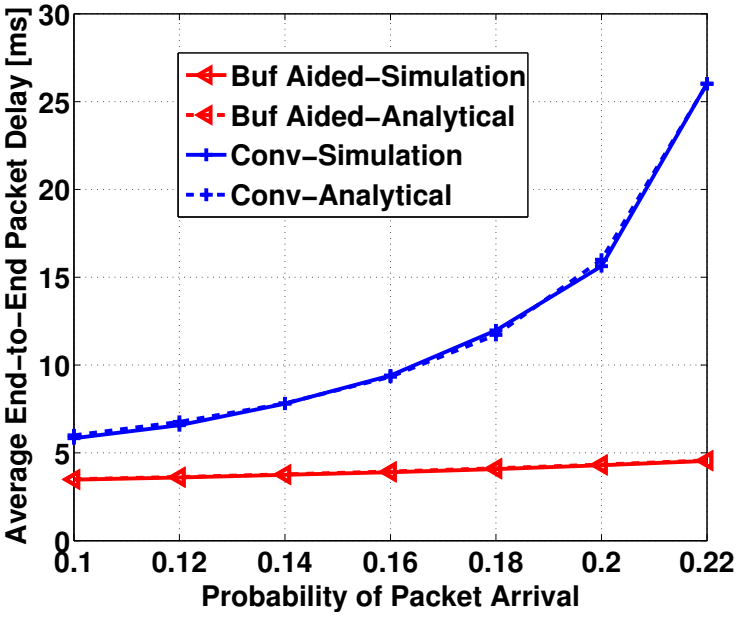
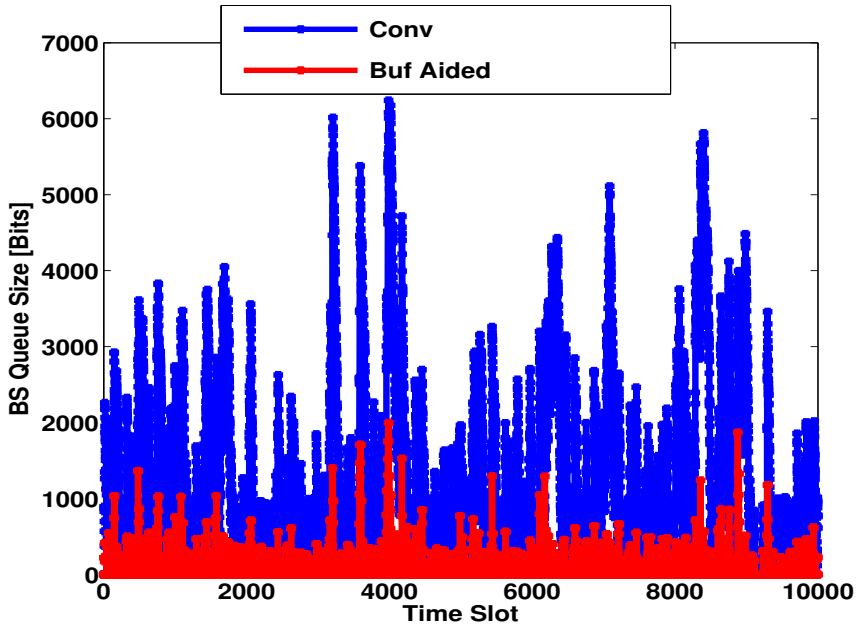


Figure 2.6: Average end-to-end packet delay in the case of Bernoulli channel distribution with $s_1 = s_2 = 0.5$

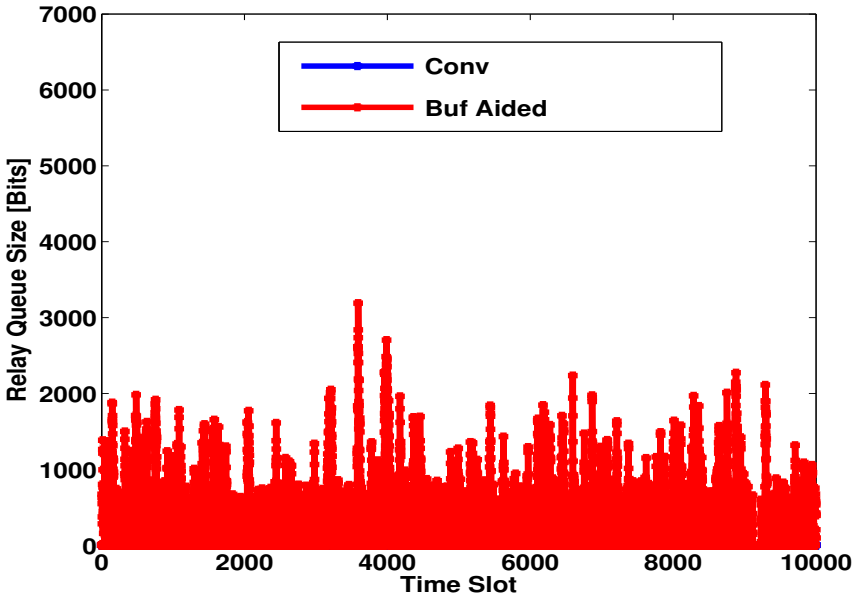
2.4.2 General Scenario

Note that the mathematical analysis presented in subsection 2.3.2 and the numerical results shown in subsection 2.4.1 are for Bernoulli data arrivals and channel conditions and provide an insight on the effect of using buffer in relay on the average end-to-end packet delay. In order to verify the discussions presented in subsection 2.3.3 for general data arrival and channel condition processes, we consider a scenario with more realistic settings. For that, the simulation parameters are selected as follows. We consider a single cell with radius 1000 m where the BS is located at the center, the relay is located at the distance of 1/2 cell radius from the BS and the user is on the cell edge (at the distance of 1/2 cell radius from the relay). BS, relay and user antenna heights are considered 15 m, 10 m and 1.5 m respectively, and the path loss model is based on [61]. The carrier frequency is 1900 MHz, the channel bandwidth is considered equal to 180 kHz and the time slot duration is set to 1 ms. The transmission power at the BS and relay is 23 dBm and the noise power spectral density is assumed -174 dBm/Hz. User data traffic is assumed Poisson with packet sizes equal to 1 kbits. It is assumed that the channel fading is flat over the system bandwidth and constant during each time slot; however, it can vary from one slot to another. For the link between the relay and user, Rayleigh channel model is used, and for the link from the BS to relay, Rician channel model is used with κ factor equal to 6 dB [62]. In the case of conventional relaying, the transmissions at the BS and relay are done in consecutive subslots. For buffer-aided relaying, we have used MW policy [54, 55, 60] to decide in an adaptive way, about the transmission in each subslot either from the BS or relay buffer.

Fig. 2.7 shows the BS and relay queue sizes over time at the data arrival rate of 50 packets/second. It is observed that with buffer-aided relaying, although there is queueing in the relay buffer, the BS queue size in each time slot is reduced significantly. This is because the BS has more flexibility to use its channel and transmit to relay when it is in



(a)



(b)

Figure 2.7: (a) BS queue size over time (b) relay queue size over time; at the arrival rate of 50 packets/slot.

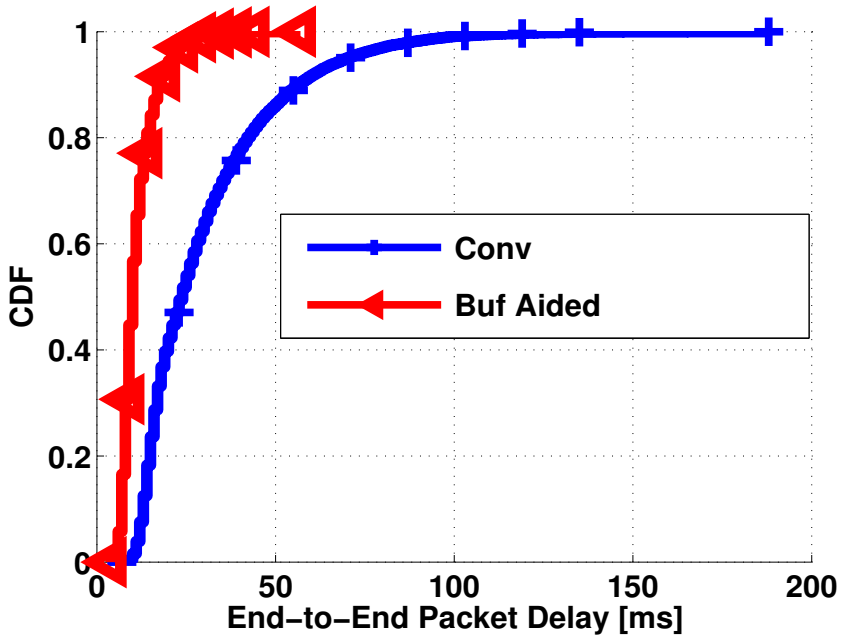
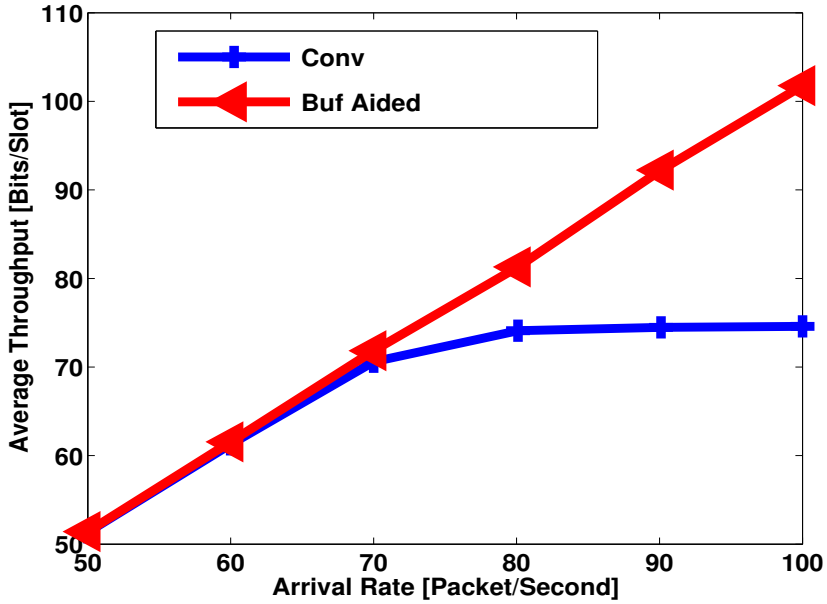


Figure 2.8: CDF of end-to-end packet delays at the arrival rate of 50 packets/slot

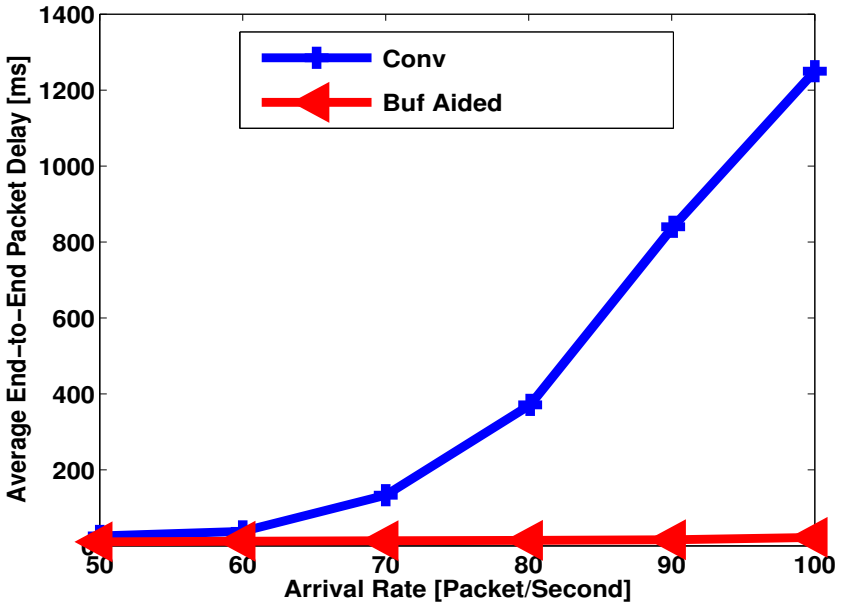
good condition, which leads to more data departures from its queue compared with the case of conventional relaying. On the other hand, having buffer at the relay allows it to queue the received data when the channel from the relay to the user is not good and transmit when it is favorable. On the whole, the channel variations are used more opportunistically and this results in the faster transfer of data from the BS to the user and lower end-to-end packet delays in buffer-aided relaying compared with conventional relaying.

The aforementioned effect can be observed through the cumulative distribution function (CDF) of the end-to-end packet delays depicted in Fig. 2.8. Fig. 2.8 indicates that even though some packets experience higher end-to-end delay in the case of buffer-aided relaying system compared with the conventional one, *average* end-to-end packet delay is less in buffer-aided relaying system. In particular, in this scenario, the average end-to-end packet delays are 11 ms and 30 ms in buffer-aided and conventional relaying systems, respectively.

Fig. 2.9 displays the effect of increase in the packet arrival rate on the throughput



(a)



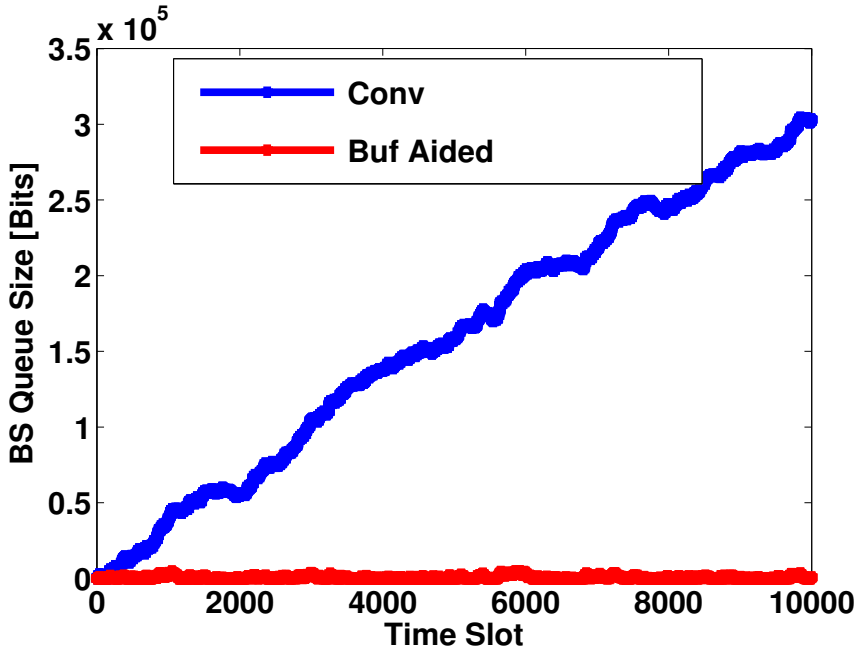
(b)

Figure 2.9: Effect of packet arrival rate at the BS on (a) average throughput in each time slot (b) average end-to-end packet delay.

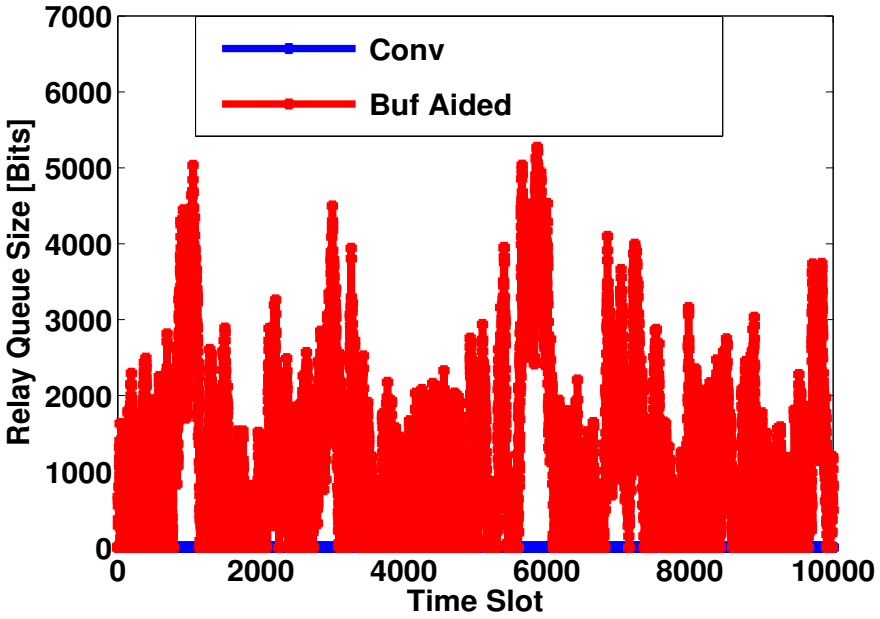
and delay performance. It is observed that, up to the arrival rate of 60 packets/second, conventional relaying is able to serve the arrived data and result in the same amount of throughput. However, after that, due to low capacity, it leads to queue instability. The effect of this is that the data departure rate of the queues are not equal to their data arrival rates and therefore, the average throughput is less than the average data arrival rate at the BS, and the packets experience large end-to-end delays. In contrast, buffer-aided relaying is able to provide average throughput equal to the average data arrival rate at the BS, in all the packet arrival rates, and therefore, leads to very low end-to-end packet delays.

In order to have a complete picture, we also present the system performance at the arrival rate of 100 packets/second. Fig. 2.10(a) shows that in conventional relaying, the BS queue grows unbounded; this is due to the low capacity of relaying channel, which is unable to serve all the arrived data. This leads to large end-to-end packet delays as depicted in Fig. 2.11. On the other hand, as shown in Fig. 2.10(b), buffer-aided relaying leads to queueing in the relay buffer, which helps to utilize the channel variations efficiently. It allows to transfer the data from the BS buffer to relay buffer and from relay buffer to user, when the corresponding channels have good conditions, and therefore, leads to low end-to-end packet delays. In particular, in this scenario, the average end-to-end packet delays are 22 ms and 1250 ms, respectively in buffer-aided and conventional relaying.

The above results confirm that using buffer at relay, improves the throughput as well as the average end-to-end packet delay in the system.



(a)



(b)

Figure 2.10: (a) BS queue size over time (b) relay queue size over time; at the arrival rate of 100 packets/slot.

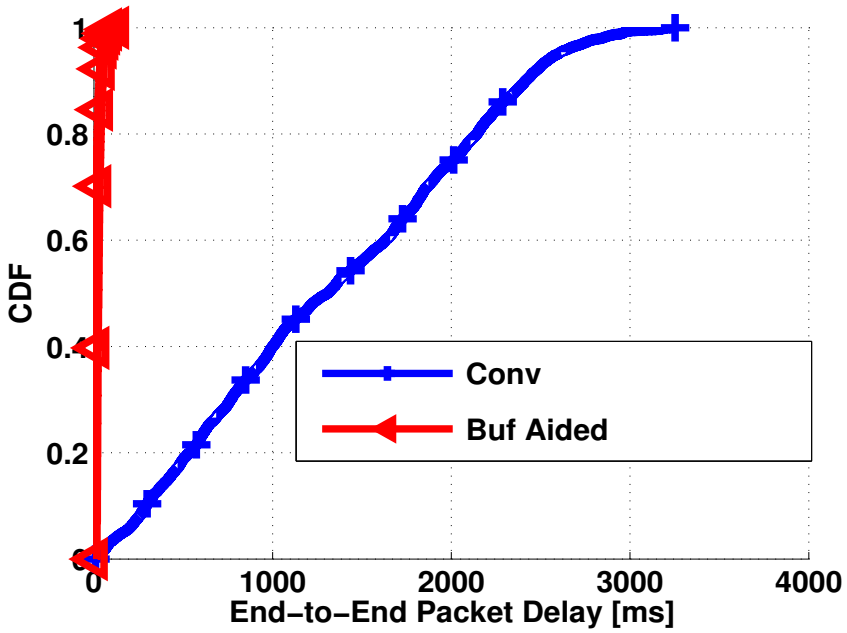


Figure 2.11: CDF of end-to-end packet delays at the arrival rate of 100 packets/slot

2.5 Summary

In this chapter, we have studied the effect of buffering at relay on the end-to-end delay performance of the system. Through the discussions about the queueing delay, we have explained the cause of delay in a simple queueing system. Based on that, we have provided an insight on the overall queueing delay in conventional and buffer-aided relaying networks. Then, through analytical investigation and intuitive generalization, we have concluded that employing buffer at the relay improves the average end-to-end packet delay. Using numerical results, we have verified our analysis and discussions, and have shown that using buffers at the relay leads to higher system throughput and lower average end-to-end packet delay.

Chapter 3

Channel-, Queue-, and Delay-Aware Resource Allocation

3.1 Introduction

In the previous chapter, we showed that exploiting buffer in relay node leads to improvements in the system throughput as well as the average end-to-end packet delay. Due to these advantages, it is expected that buffer-aided relay networks will attract a lot of attention for using in cellular networks. Recently, there has been some research on resource allocation in such networks, and novel algorithms for scheduling and subchannel allocation have been proposed [47–49]. However, it is still needed to investigate in such scenarios, the QoS provisioning for delay-sensitive services with strict delay requirements, together with delay-tolerant users having average throughput requirements. There has been QoS-aware algorithms already proposed for prompt relays [23–25]; however, for buffer-aided relays, new algorithms need to be designed to take into account the queuing delay in the relays in meeting the deadline of the delay-sensitive packets. Also the works such as [23–25] have considered only one of the above QoS requirements, and therefore, the scenarios where users with different QoS requirements are present in the network necessitates more investigation. As it will be discussed in Section 3.3, in order to meet the deadlines of delay-sensitive packets, minimum transmission rates need to be considered over the links; however, depending on the routing path from the BS to end users, the links with minimum

rate requirements would be different. Due to this and also depending on the method to compute these minimum rate requirements, different constraints would be imposed on the network in each scheduling period. All of these will affect the capacity to provide average throughput guarantees for delay-tolerant users. In this chapter, our goal is to investigate the above issues and propose potential policies for addressing them.

In particular, we study QoS-aware routing and subchannel allocation in the downlink of time-slotted OFDMA networks enhanced with buffering relays. The main contributions of our work are summarized as follows:

- We consider a heterogeneous service environment, where the goal is to meet the packet deadlines of delay-sensitive users and at the same time maintain the queue stability in the system, to ensure that the average throughput seen by each delay-tolerant user matches the average data arrival rate at the BS.

- We present a framework for Time Domain Scheduling (TDS) and Frequency Domain Resource Allocation (FDRA), through which we show that the IRAP formulation for meeting the stated QoS objectives is itself a challenge. Based on this framework, we identify the issues that need to be addressed in formulating IRAP and the design of scheduling and routing algorithms.

- We propose novel CQDA policies to define the parameters needed for IRAP formulation in each time slot. Specifically, these policies take several steps to determine the utility function and minimum rate constraints over links, and at the last step, they use an iterative algorithm to solve the problem they have formulated.

- Using extensive simulations, we evaluate the performance of the proposed policies and show that they are able to utilize the system resources well to reach the objectives stated.

The rest of this chapter is organized as follows. Section 3.2 describes the system model for a buffer-aided relay-enhanced OFDMA network. In Section 3.3, we outline the resource allocation objectives and discuss the challenges for formulating IRAP. Section 3.4 presents the proposed CQDA policies and practical considerations. Simulation results are provided in Section 3.5, and Section 3.6 gives the conclusions.

3.2 System Model

As shown in Fig. 3.1, we consider the downlink of a time-slotted OFDMA system in a single cell with K users, M relays and N subchannels. Each subchannel consists of several subcarriers to reduce the overhead of signaling about the channel conditions and assignments. Users, relays and subchannels are indexed respectively by $k \in \{1, ..., K\}$, $m \in \{1, ..., M\}$ and $n \in \{1, ..., N\}$. Users who are close to the BS and have good link quality are served only by the BS. We refer to these users as "close" users and group them in the set $\breve{\mathcal{K}}$. On the other hand, the users far from the BS are also assigned to one or more relays, from which they can receive high signal strength and, therefore, high data rate. We refer to these users as "far" users and group them in the set $\widehat{\mathcal{K}}$. These terms are used instead of "direct" and "relayed" because we assume that all the users have a direct connection to the BS, whereas "relayed user" might be misinterpreted as a user that has no direct connection to the BS. Further, we have used the term "direct" in the following to specify the link type and it might cause ambiguity if the term "direct user" was used. We use $\mathcal{K}_m$ to denote the set of users that can receive service from $m = 0, 1, ..., M$, where $m = 0$ indicates the BS and therefore, $\mathcal{K}_0 = \{1, ..., K\}$. Similarly, we use $\mathcal{M}_k$, $k = \{1, ..., K\}$, to refer to the serving nodes of the user k, which could be one or more relays and/or the BS; e.g. $\mathcal{M}_2 = \{0, 1, 6\}$ indicates that user 2 is able to receive data from the BS and relays 1 and 6. We assume that the sets of users and relays defined above are determined at the

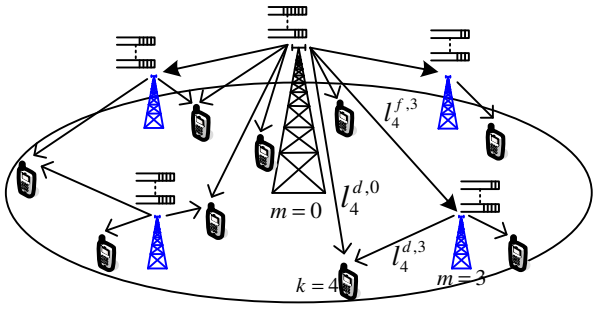


Figure 3.1: System model

beginning of users' connections to the network and remain unchanged.

The transmission links are uniquely classified into two types. First are the links between a serving node m, $m = 0, ..., M$, and the users that can receive data from it, i.e., $k \in \mathcal{K}_m$. The variables that are defined for this type of links have the superscript d beside the letter m, which indicates that the link is a "direct" one from node m to the user; e.g., $l_k^{d,m}$ denotes the direct link from node m to user k and $e_{kn}^{d,m}$ indicates the channel gain of this link on subchannel n. The second type of links are the far users' feeder links between the BS and relays, for which we use the superscript f to indicate that the link is a "feeder" link between the BS and relay m, $m = 1, ..., M$. It is important to note that in this case, the letter m does not show the transmitting node but the receiving one; e.g., $l_k^{f,m}$ denotes the feeder link of user k from the BS to relay m and $e_{kn}^{f,m}$ denotes the channel gain of this link on subchannel n. We assume that the channel gains of the links vary over time and frequency, but remain constant on each subchannel during a time slot.

It is assumed that the BS and relays are equipped with buffers; the BS buffers are fed by exogenous packet arrival processes and served by transmissions from the BS to users or relays. On the other hand, buffers in relays are fed by the BS transmissions on feeder links and are served by transmissions to their users. We use $Q_k^m(t)$, $k \in \mathcal{K}_m$, to denote the number of bits queued in the buffer of user k in node m, $m = 0, ..., M$, at time slot t. Also we assume that the relays are quasi full-duplex and have the ability to receive and

transmit at the same time slot but on different subchannels. As a result, in each time slot some subchannels can be used for transmissions on direct links (from the BS or relays to users) and some for feeder links (from the BS to relays). However, there is no frequency reuse inside the cell and each subchannel can only be assigned to one link[2].

For simplicity of the system model, it is assumed that a fixed power p is used for transmission on any subchannel by any node and this power is equally distributed on the subcarriers forming the subchannel. Assuming that M-ary quadrature amplitude modulation (QAM) is used for transmissions, the achievable transmission rate in each time slot between node m and user k on subchannel n can be computed as follows [63]:

$$r_{kn}^{d,m} = BT \log_2 \left(1 + \frac{p e_{kn}^{d,m}}{\Gamma_k B \nu_0}\right), \tag{3.1}$$

where B and T are the bandwidth of a subchannel and the time slot duration, respectively. ν_0 denotes the noise spectral density and Γ_k is the signal-to-noise ratio (SNR) gap due to the limited number of coding and modulation schemes, which is related to bit error rate (BER) of user k (BER_k) through equation $\Gamma_k = -\frac{\ln(5BER_k)}{1.5}$ [63]. In a similar way, based on $e_{kn}^{f,m}$ and Γ_k, we can define $r_{kn}^{f,m}$ as the achievable transmission rate of the feeder link of user k between the BS and relay m on subchannel n. Note that due to different BER requirements of the users, the achievable rate on the feeder link from the BS to relay m for two different users can be different. Using (3.1), the transmitted number of bits in each slot on the link from node m to user k is equal to:

$$r_k^{d,m} = \sum_{n=1}^{N} x_{kn}^{d,m} r_{kn}^{d,m} \tag{3.2}$$

where $x_{kn}^{d,m}$ denotes the subchannel allocation indicator, which equals one if subchannel n

[2]Although in cooperative transmissions, the BS and relays can transmit simultaneously on the same subchannel, for simplicity we do not consider this possibility.

is used for transmission on the link from m to user k, or zero otherwise. Similarly $r_k^{f,m}$ can be computed based on $x_{kn}^{f,m}$ and $r_{kn}^{f,m}$. Note that in the following, depending on the context, we may use e_n^l, r_n^l, x_n^l to denote respectively $e_{kn}^{d,m}$, $r_{kn}^{d,m}$, $x_{kn}^{d,m}$ when $l \in \{l_k^{d,m}\}$, or $e_{kn}^{f,m}$, $r_{kn}^{f,m}$, $x_{kn}^{f,m}$ when $l \in \{l_k^{f,m}\}$.

We consider a heterogeneous traffic scenario, where some users have delay-tolerant traffic and others have delay-sensitive one. A delay-tolerant traffic does not have strict deadlines for its packets, but requires an average throughput guarantee to make sure its packets are delivered to the destination with a finite average delay. On the other hand, the packets of delay-sensitive traffic have a maximum allowed delay, after which the packets become expired and therefore, dropped from the corresponding buffers. We assume that the packet arrivals of all traffic streams at the BS are stationary and ergodic processes and have finite mean and variance. We use $\mathcal{K}^a$ and $\mathcal{K}^b$ to denote the set of delay-sensitive and delay-tolerant users, respectively. For any $k \in \mathcal{K}^a$, $\hat{W}_k$ indicates the maximum allowed delay for its packets. It is assumed that each packet of delay-sensitive traffic is tagged with a number that is updated every time slot and shows the packet's delay since its arrival in the BS queue. We use $I_k^m(t)$, W_{ki}^m, and L_{ki}^m to denote, respectively, the number of packets at time slot t, the delay, and the size of i-th packet in the queue of user k in node m. The packets are indexed in the same order as their arrival; therefore, a packet with index 1 is also referred as the Head of Line (HoL) packet. It is clear that $Q_k^m(t) = \sum_{i=1}^{I_k^m(t)} L_{ki}^m$. We assume that the buffers have infinite capacity and therefore, no packet drop occurs due to buffer overflow. As a result, the number of packet drops is always zero for delay-tolerant users.

We assume that a centralized scheduler in the BS has perfect knowledge about the queue sizes and packet delays in the BS and the relay buffers as well as channel states of all the links, based on which it decides about the allocation of subchannels.

3.3 Quality-Of-Service-Aware Resource Allocation Problem

In this section, first we explain the main objectives and then discuss the challenges in formulating the corresponding IRAP.

3.3.1 The Main Objectives

As explained in the previous section, we consider both delay-tolerant and delay-sensitive users. Delay-sensitive users have strict delay constraints for their packets; therefore, one of the objectives of the scheduler is to allocate enough resources to these users in order to make sure that their packets arrive in the receiver on time, or at least to keep the number of expired packets low. On the other hand, although the delay-tolerant users do not have strict deadlines for their packets, the scheduler aims at providing them with an average throughput equal to the average data arrival rates at their queues in the BS. This is equivalent to guaranteeing that their packets would experience finite average queueing delays in the buffers of the BS and relays [55]. Therefore, the other objective is to make sure that the queues belonging to delay-tolerant users are stable, i.e., their sizes remain bounded[3]. Based on the above, the main objectives can be summarized as providing the following, subject to the system's physical constraints (which was mentioned in the previous section, i.e., the limitations on the system resources and their usage):

$$1)\ W_{ki}^{m} \leq \hat{W}_k, m \in \{0, ..., M\}, k \in \mathcal{K}_m \cap \mathcal{K}^a, \forall t, i \in \{1, ..., I_k^m(t)\}, t = 0, 1, ... \tag{3.3a}$$

$$2)\ \lim_{\tau \to \infty} \sup \frac{1}{\tau} \sum_{t=0}^{\tau-1} E[Q_k^m(t)] < \infty, m \in \{0, ..., M\}, k \in \mathcal{K}_m \cap \mathcal{K}^b \tag{3.3b}$$

[3]Note that the queues of delay-sensitive users are always stable due to drops of expired packets.

where (3.3a) means that the delay of any packet of a delay-sensitive user k in its queues in the BS and any serving relay should be less than the threshold $\hat{W}_k$, and (3.3b) is based on the definition of the strong stability [55] and states its satisfaction for the queues of delay-tolerant users in the BS and relays. In this equation, E[.] indicates expected value. Note the underlying assumption that an admission control entity in a higher layer exists to make decisions about admitting users, while taking their QoS requirements into account. Based on those, it defines the above objectives and then the scheduler aims at provisioning them. Therefore, if for example an admitted user's data arrival rate is higher than another, the scheduler will need to provide more service to it to keep its queues stable; or if a deterministic delay bound is also desired for admitted delay-tolerant users, the objective (3.3a) will be applied on them instead of (3.3b) and the scheduler will need to treat them as delay-sensitive users by tagging their packets with their delay. Also, since our goal is to investigate the service provisioning for users with QoS requirements, we do not consider BE users and the subject of fair service provisioning for them. These types of users can be served based on PF scheduling, by considering dedicated subchannels for them or by the resources left unused after allocating the subchannels to the users with QoS requirements.

3.3.2 Challenges of IRAP Formulation

The objectives stated above are to be satisfied in the long term. For this purpose, it is needed to translate these objectives into IRAPs over the time slots to come. In other words, a dynamic scheduling mechanism should monitor the channel states, queue sizes and the delays of the packets in each time slot, and based on their instantaneous values, it should first formulate the specific problem to be addressed in that time slot and then use an algorithm to solve it.

Regarding the second objective, it has been shown in [54, 55] that in a multi-hop wireless

network, applying the well-known MW policy can make all the queues stable, as long as it is feasible to do so by any other algorithm. MW reaches this objective by considering the IRAP as maximizing the sum of the weighted rates of the links in each time slot, subject to the constraints of the system. The weight of a link is considered equal to the difference of the corresponding queue sizes in the beginning and the end points of that link[4].

However, in our system the challenge is how to define and include the constraints that the delay-sensitive users impose on the network (through the first objective), while reserving enough capacity for stabilizing the queues of delay-tolerant users. In the literature on single-hop networks, in order to meet the deadlines of delay-sensitive packets and maximize the throughput of delay-tolerant users, resource allocation is usually divided into time domain and frequency domain stages: A time domain scheduler determines a minimum transmission rate for delay-sensitive users in each slot, and then a frequency domain scheduler allocates the subchannels to first satisfy the minimum rate requirements and then to maximize the throughput of delay-tolerant users [64]. However those works do not consider the queue stability of the delay-tolerant users and also it is not clear how to extend their policies to two-hop networks.

To clarify the above mentioned, we present the QoS-aware cross layer scheduling and resource allocation with the following framework for formulating an optimization problem by the BS, and discuss the challenges. In fact, for translating the objectives in (3.3a) and (3.3b) into IRAPs over time slots, BS considers this framework in each time slot:

1. Determine $\mathcal{K}_m^o \subseteq \mathcal{K}_m$, $\forall m$, and $R_k^{y,m}$, $y = d, m = 0, ..., M, k \in \mathcal{K}_m$, $y = f, m = 1, ..., M, k \in \mathcal{K}_m$, for the following FDRA problem, in a way to meet the packet deadlines and provide queue stability. $\mathcal{K}_m^o$ is the set of users to be considered in the utility function (3.4a) and $R_k^{y,m}$ is the minimum rate constraint for the transmission

[4]It is assumed that the data packets exit the system when they reach their users and therefore, the queue sizes in user nodes are zero.

on link $l_k^{y,m}$ (from the queue in the beginning point of the link), also denoted by $R^l, l \in \{l_k^{y,m}\}$ for simplicity.

2. Solve the instantaneous frequency domain resource allocation problem stated in (3.4).

$$\max_{\boldsymbol{x}} \sum_{m=0}^{M} \sum_{k \in \mathcal{K}_m^o} \sum_{n \in \mathcal{N}} x_{kn}^{d,m} w_k^{d,m} r_{kn}^{d,m} + \sum_{m=1}^{M} \sum_{k \in \mathcal{K}_m^o} \sum_{n \in \mathcal{N}} x_{kn}^{f,m} w_k^{f,m} r_{kn}^{f,m}, \tag{3.4a}$$

$$\text{s.t.} \quad C1: \sum_{n \in \mathcal{N}} x_{kn}^{y,m} r_{kn}^{y,m} \geq R_k^{y,m}, \tag{3.4b}$$

$$C2: \sum_{n \in \mathcal{N}} [x_{kn}^{d,0} r_{kn}^{d,0} + \sum_{m \in \mathcal{M}_k} x_{kn}^{f,m} r_{kn}^{f,m}] \leq Q_k^0, \forall k, \tag{3.4c}$$

$$C3: \sum_{n \in \mathcal{N}} x_{kn}^{d,m} r_{kn}^{d,m} \leq Q_k^m, m \in \{1, ..., M\}, k \in \mathcal{K}_m, \tag{3.4d}$$

$$C4: \sum_{m=0}^{M} \sum_{k \in \mathcal{K}_m} x_{kn}^{d,m} + \sum_{m=1}^{M} \sum_{k \in \mathcal{K}_m} x_{kn}^{f,m} \leq 1, \forall n, \tag{3.4e}$$

$$C5: x_{kn}^{d,m} \in \{0, 1\}, \forall n, \forall m \in \{0, ..., M\}, k \in \mathcal{K}_m, \tag{3.4f}$$

$$C6: x_{kn}^{f,m} \in \{0, 1\}, \forall k \in \mathcal{K}_m,\ x_{kn}^{f,m} = 0, \forall k \notin \mathcal{K}_m, \forall m \in \{1, ..., M\}, \forall n \tag{3.4g}$$

In (3.4), $\boldsymbol{x} = \{x_{kn}^{d,m}\} \cup \{x_{kn}^{f,m}\}$ is the set of all indicator variables. We consider the weights equal to $w_k^{d,m} = \alpha_k^{d,m} \beta_k Q_k^m$, $w_k^{f,m} = \alpha_k^{f,m} \beta_k (Q_k^0 - Q_k^m)$ to provide stability (this is achieved by considering the queue backlogs) and also prioritize users' links based on their locations and QoS requirements (this is considered by $\alpha_k^{y,m} \beta_k$ coefficients). $C1$ states the minimum rate constraints that need to be determined in order to meet the packet deadlines, and $C2$ and $C3$ state the maximum number of bits that can be actually sent from the queues of the BS and relays, respectively. $C4$ ensures that each subchannel is allocated exclusively to a single link and constraint $C5$ and $C6$ define the possible values for channel allocation indicators of direct and feeder links, respectively.

It is seen that for the formulation of the IRAP, its parameters need to be determined first. We note that the utility function (3.4a) has been chosen based on the MW policy (also known as backpressure routing method) [54]. This way, subchannels would be allocated to the links that have good channel conditions and/or large differential queue backlogs at the two ends. As a result, data packets will automatically be routed through the links that have good channel conditions and/or less queue congestion in the next hop. Also, if a queue does not get serviced for a while (e.g., when the channel conditions of its serving links are not good) and its size increases, its weight will grow large; this will lead to providing service to it, which in turn will reduce its size. The issue here is how the set of users $\mathcal{K}_m^o$, $\forall m$, in the utility function (3.4a), should be determined. Should they include only delay-tolerant users or delay-sensitive users as well?

The other issue is how the minimum rate requirements of a delay-sensitive user should be defined. Should they be defined in every time slot or just in some specific time slots? Also, we note that the queues of users exist both in the BS and relays and it is not known beforehand how long the delay of the packets will be in each queue. Therefore, one other important thing that should be decided is the route of the delay-sensitive packets for far users. In other words, should these packets be transmitted directly from the BS or through relays? The decision about this will determine on which links the minimum rate constraints (3.4b) should be imposed. The other challenge is the decision for the delay budget division between the BS and relays for the delay-sensitive users, in case their packets are routed through the relays. This is important for computing the values of the minimum transmission rate requirements mentioned above.

All the above issues are inter-related and will affect the IRAP formulation. To the best of our knowledge, due to the individual delay requirements for delay-sensitive packets, stochastic nature of data arrivals and channel variations as well as complexity of network

architecture, design and analysis of an optimal method is highly intractable. Therefore, in the next section, we will propose different suboptimal policies for addressing the stated issues in IRAP formulation and solving the resulting IRAPs. Then, in Section 3.5, we will use simulations to verify the performance of the proposed policies.

3.4 Quality-Of-Service-Aware Cross Layer Scheduling and Resource Allocation

In this section, we propose novel QoS-aware resource allocation policies for addressing the challenges mentioned in the previous section. We present in detail the steps that these policies take to formulate and solve the IRAP, and discuss the practical considerations for them.

3.4.1 CQDA Policies

CQDA policies first decide about the needed parameters for the utility function and minimum rate constraints, to formulate the IRAP in each time slot; due to their different approaches, each of these policies results in different parameters and, therefore, a different instance of the problem (3.4). Then, these policies use a similar iterative subchannel allocation algorithm to solve their formulated problem. Fig. 3.2 shows the main steps and substeps that these policies take in every time slot. Based on the QCSI and packet delays in each time slot, the BS can use CQDA policies to decide about the routing and scheduling of the users' packets, for reaching the main objectives stated in (3.3).

The approaches that the proposed policies have for parameter decision and IRAP formulation can be summarized as follows.

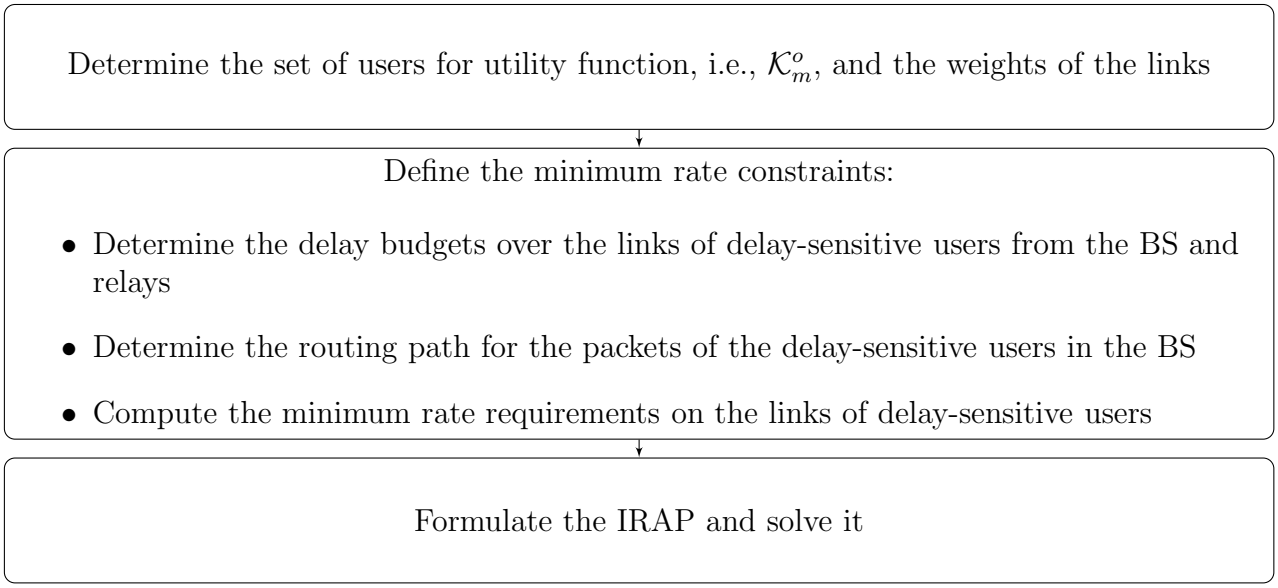


Figure 3.2: Flowchart of the IRAP formulation

a) Policy 1: Separate Utility and Minimum Rate (SUMR) Definitions. In this policy, the utility function in every time slot is defined only based on the channel and queue states of the delay-tolerant users in that time slot. On the other hand, the minimum rate requirements of the delay-sensitive users are computed and applied whenever there are packets in their queues in the BS or relays. In fact, the problem formulation approach of this policy targets the provisioning of (3.3a) and (3.3b) separately: by guaranteeing the service of delay-sensitive users through the minimum rate constraints only and the service of the delay-tolerant users through the utility maximization.

b) Policy 2: Joint Utility and Minimum Rate (JUMR) Definitions. This policy defines the utility function in each time slot based on the channel and queue states of all the users in that time slot, but with higher priorities for delay-sensitive users. In this policy, the minimum rate requirements are defined and considered only when there are packets that have reached the thresholds determined in step 2, and have not yet been transmitted under the effect of utility maximization in time slots before. The approach this

policy uses for problem formulation in fact aims at provisioning of (3.3a) and (3.3b) jointly: by guaranteeing the service of all the users through the utility maximization, except when the delay-sensitive packets are due, in which case the minimum rate constraints are also applied.

In the following we explain in detail all the steps that these policies take to formulate the IRAP in each time slot, as well as the specific differences that they have in each step.

Step 1) Determine the set of users for utility function and the weights of the links.

a) In the SUMR policy, $\mathcal{K}_m^o = \mathcal{K}_m \cap \mathcal{K}^b, \forall m$. This way, the links considered in the utility function will be all that belong to delay-tolerant users. For the α coefficients in the weights of the links in (3.4a), this policy considers the following:

$$\alpha_k^{y,m} = \begin{cases} \frac{1}{\overline{r}_k^{d,m}}, & y = d, m = 0, \forall k \\ \frac{1}{\frac{1}{2}\min(\overline{r}_k^{f,m}, \overline{r}_k^{d,m})}, & y \in \{d, f\}, m = 1, ..., M, k \in \mathcal{K}_m \end{cases} \tag{3.5}$$

where $\overline{r}_k^{y,m}$ is the average achievable rate for the link $l_k^{y,m}$ on a single subchannel (considering the mean channel gain, i.e., without the effect of small scale fading). By adjusting the weights of the links through the above values for α's, we try to compensate for the effect of distance on transmission rates and provide some fairness. For example if user k has a larger distance from the BS and the channel condition for direct transmissions from the BS to it is low on average, it will have a larger $\alpha_k^{d,0}$ and, therefore, will be selected more often to get service. Note that according to the second line in (3.5), the feeder link from the BS to relay and the direct link from that relay to the user are assigned the same value of α. This value is computed based on the average achievable rate of the link that has lower channel conditions on average, as it is the one that limits the overall two-hop transmission rate; the coefficient $\frac{1}{2}$ is

due to the fact that the wireless resources are used twice in two-hop transmissions. SUMR considers β coefficients equal to one for all the delay-tolerant users, as they belong to the same traffic class[5].

b) In the JUMR policy, $\mathcal{K}_m^o = \mathcal{K}_m, \forall m$. Since the minimum rate requirements in this policy are defined only when the HoL packets reach the determined threshold over their routing link (clarified in the next step), the utility is thus defined also for the queues of delay-sensitive users to be served enough before reaching the packet deadlines. Furthermore, in order to prevent the cases in which the HoL packets in several queues reach a deadline and require minimum rates that might not be met altogether, the delay-sensitive users' queues should be provided with higher service rates than those of delay-tolerant users. For this purpose, other than setting α coefficients as in (3.5), a higher priority is also considered for them, which is applied on the weights of their links through setting their β coefficients to a value several times larger than 1. In our simulations, we have found that using the inverse of the maximum allowed delay of each user, i.e., $\beta_k = \frac{1}{\hat{W}_k}, \forall k \in \mathcal{K}^a$, gives a good performance. This is because $\hat{W}_k$ is usually in the order of about one hundred millisecond, the inverse of which gives delay-sensitive users more than five times priority over delay-tolerant users.

Step 2) Define the minimum rate constraints. For defining the minimum rate constraints, each of the policies need to decide about the delay thresholds for transmissions over the links, the routing paths and the values of minimum rate requirements for delay-sensitive packets. Therefore, this step can be further split into the substeps explained in the following.

Substep 2-1) Determine the delay budgets over the links of delay-sensitive users from the BS and relays

[5]The β coefficients can be considered differently, in cases that users are also classified based on other metrics than the traffic classes, like different service plans.

In this substep, the goal is to consider a rule for delay budget division of delay-sensitive packets on the two hops. This is to make sure that in the case of routing through relays, the delay-sensitive packets will have enough time left for transmissions from the relay to their users before expiring. Based on this, in the next substeps, the policies will decide about the routing and the values of minimum rate requirements.

For this purpose, we consider a delay threshold corresponding to the links of the users: define $D_k^{y,m}$ as the delay threshold for the transmissions on link $l_k^{y,m}$, from the queue in the beginning point of the link. Also, depending on the context, we represent this by D^l. Due to the fact that the packets are tagged with their delay since the arrival time in the BS (not the arrival time in the current queue, which may be in a relay), the thresholds for direct transmissions from the BS and relay queues are thus equal to the packets' thresholds, i.e., $D_k^{d,m} = \hat{W}_k$ for $m = 0, ..., M, k \in \mathcal{K}_m \cap \mathcal{K}^a$. Consequently, the delay budget division can be specified by defining a delay threshold for the feeder links between the BS and relays, as explained in the following.

a) SUMR shares the delay budget between the BS and relays in two-hop transmissions based on the average transmission rates on the link from the BS to relay and the link from relay to user. Specifically, we have

$$D_k^{f,m} = round(\hat{W}_k \frac{\overline{r}_k^{d,m}}{\overline{r}_k^{f,m} + \overline{r}_k^{d,m}}), m = 1, ..., M, k \in \mathcal{K}_m \cap \mathcal{K}^a. \tag{3.6}$$

The intuition behind (3.6) is the fact that the transmission time of a packet is proportional to the inverse of its transmission rate; if the length of packet i in the queue of the user k in the BS is L_{ki}^0, an estimate of the average time for its (continuous) transmission from the BS to relay $m \in \mathcal{M}_k$ and from relay m to the user k, on a single (dedicated) subchannel, would be $\tau_1 = \frac{L_{ki}^0}{\overline{r}_k^{f,m}}$ and $\tau_2 = \frac{L_{ki}^0}{\overline{r}_k^{d,m}}$, respectively. Noting that $\frac{\tau_1}{\tau_1+\tau_2} = \frac{\overline{r}_k^{d,m}}{\overline{r}_k^{f,m}+\overline{r}_k^{d,m}}$, SUMR uses this estimated ratio for computing the delay

budget share of the feeder link.

b) JUMR considers $D_k^{f,m} = \hat{W}_k - 1,\ m = 1, ..., M, k \in \mathcal{K}_m \cap \mathcal{K}^a$ for transmissions on feeder links. This is due to the fact that JUMR tries to serve the queues of the delay-sensitive users mostly through the utility maximization (as described in the previous step), and therefore, the routing path and the transmission rates are automatically specified by subchannel allocation to the links (this will be clarified later). Consequently, it does not need to work based on the delay budget division, except in the last moments of packet deadline; i.e., based on the above setting for $D_k^{f,m}$, JUMR starts to decide about the routing and the minimum rates for the queues of far delay-sensitive users in the BS, one slot before the deadline. This way, if the two-hop transmission is decided in the next substep, there will be time for that: one time slot for BS to relay transmission and one time slot for relay to user transmission.

Substep 2-2) Determine the routing path for the packets of the delay-sensitive users in the BS[6]

The main goal in this substep is to decide if the packets in the queues of far delay-sensitive users in the BS should be transmitted through the direct or feeder links from the BS. Based on this decision, the minimum rate requirements of those packets are then imposed on the direct or feeder links from the BS. For the packets in the queues of close delay-sensitive users in the BS and the queues of far delay-sensitive users in the relays, there is no need for routing and the transmissions are done on the corresponding direct links between the BS/relay and those users. We show this by

$$s_k^m = l_k^{d,m}, \quad m = 0, k \in \breve{\mathcal{K}} \cap \mathcal{K}^a; \quad m = 1, ..., M, k \in \mathcal{K}_m \cap \mathcal{K}^a \tag{3.7}$$

[6]Note that for the delay-tolerant users, packets are always routed automatically as a result of the subchannel allocations based on the utility maximization.

where s_k^m indicates the link over which a minimum rate will be determined for transmissions from the queue of user k in node m. In order to propose a routing method for the packets in the queues of far delay-sensitive users in the BS, we define the following routing metric:

$$\rho_k^m = \frac{\overline{r}_k^{d,m}}{\overline{W}_k^m},\ k \in \widehat{\mathcal{K}} \cap \mathcal{K}^a, m \in \mathcal{M}_k \tag{3.8}$$

where $\overline{W}_k^m$ is the average delay of the data bits present in the queue of user k in node m in the current time slot, which we define as

$$\overline{W}_k^m = \frac{\sum_{i=1}^{I_k^m(t)} L_{ki}^m W_{ki}^m}{\sum_{i=1}^{I_k^m(t)} L_{ki}^m}, \quad k \in \widehat{\mathcal{K}} \cap \mathcal{K}^a, m \in \mathcal{M}_k \tag{3.9}$$

Based on this, the routing algorithm of each policy is presented in the sequel:

a) SUMR performs the routing of the far delay-sensitive users in every time slot, based on their packet delays in the BS queues, as follows:

- If $\hat{W}_k - W_{k1}^0 = 0$, $k \in \widehat{\mathcal{K}} \cap \mathcal{K}^a$, consider the direct transmission from the BS and set $s_k^0 = l_k^{d,0}$.
- If $\hat{W}_k - W_{k1}^0 > 0$, $k \in \widehat{\mathcal{K}} \cap \mathcal{K}^a$, find $\hat{m} = \arg\max_m \rho_k^m$. If $\hat{m} = 0$ consider the direct transmission from the BS and set $s_k^0 = l_k^{d,0}$; otherwise, consider forwarding to relay $\hat{m}$ and set $s_k^0 = l_k^{f,\hat{m}}$.

The logic behind the routing metric and algorithm is as follows. If the HoL packet in the queue of user k in the BS has not been sent yet and has reached its deadline (which might happen due to high load of the network or bad channel conditions of the user links), there is no point in using the routing metric and forwarding it to any relay, because if this is done, it will be dropped from the relay in the next slot. Therefore, the only chance to transmit the packet is the current time slot and

through the direct transmission from the BS. On the other hand, if there is still time to transmit the packet, then based on the routing metric, every path has a chance to be selected. The routing metric takes into account the effect of average transmission rate as well as the queuing delay and gives higher priority to the serving nodes that have larger average transmission rates and/or lower average delay.

b) In the JUMR policy, except for the cases that the packets of far delay-sensitive users have reached their corresponding delay thresholds on the feeder links, they will be routed in a manner similar to the delay-tolerant users and based on the utility maximization (e.g., when a subchannel is allocated to a feeder link, the packet is routed automatically through the corresponding relay); only when $\hat{W}_k - W_{k1}^0 = 1$, $k \in \widehat{\mathcal{K}} \cap \mathcal{K}^a$, JUMR uses the routing metric and follows the same method as mentioned above for SUMR.

Subtep 2-3) Compute the minimum rate requirements on the links of delay-sensitive users

a) In the SUMR policy, whenever there are packets in the queues of delay-sensitive users in any serving node, a minimum rate constraint is considered on the links selected (in previous substep) to serve them. This minimum rate requirement is dynamically computed in every time slot, based on the delay threshold corresponding to the selected transmission link (i.e., feeder or direct), and the sizes and delays of the packets in those queues. The details are as follows.

- If $\hat{W}_k - W_{k1}^0 = 0$, $k \in \widehat{\mathcal{K}} \cap \mathcal{K}^a$, use the size of the HoL packet as the minimum rate requirement on the direct link from the BS to that user, i.e., $R^{s_k^0} = L_{k1}^0$.
- For the transmissions on the other links selected to serve the delay-sensitive

users' queues, use the following equation:

$$R^{s_k^m} = \max_{i \in \{1,\dots,I_k^m\}} z_{k,c(i)}^m, \text{ where } z_{k,c(i)}^m = \frac{\sum_{j=1}^{i} L_{kj}^m}{(D^{s_k^m} - W_{ki}^m + 1)} \tag{3.10}$$

The given equation is based on the one proposed in [65] for guaranteeing the packet delays with low energy consumption in a single channel system; in our system since the power is fixed over the subchannels, this will lead to a low number of needed subchannels. In equation (3.10), $z_{k,c(i)}^m$ is the fixed transmission rate that can serve packets 1 to i (totally with size $\sum_{j=1}^{i} L_{kj}^m$ bits) in the queue of user k in node m over the time left until the deadline of the i-th packet on the selected link (i.e., $D^{s_k^m} - W_{ki}^m + 1$ time slots). Taking the maximum from the $z_{k,c(i)}^m$ for all the packets in that queue ensures that each individual packet gets transmitted before its deadline.

b) In the JUMR policy, similar to SUMR, if there is a packet of far delay-sensitive users in the BS, the deadline of which has arrived, $R^{s_k^0} = L_{k1}^0$; For the transmissions on any other link (specified in substep 2-2), the following equation is used:

$$R^{s_k^m} = \begin{cases} L_{k1}^m, & \text{if } D^{s_k^m} - W_{k1}^m = 0 \\ 0, & \text{otherwise} \end{cases} \tag{3.11}$$

Step 3) Formulate the IRAP and solve it. After determining the values of $\mathcal{K}_m^o$'s, w's, and R^l's through the aforementioned steps in each time slot, an instance of the problem (3.4) is obtained, which is a mixed integer nonlinear programming problem, which needs an exhaustive search to get the optimal solution and may be prohibitive in the cost of computation. Instead, we propose an iterative algorithm to allocate subchannels to the links of the users. The detailed procedure is shown in the FDRA algorithm. This algorithm

Algorithm 3.1 FDRA

Initialize $\mathcal{N} = \{1, ..., N\}$ and $q_k^m = Q_k^m, m \in \{0, ..., M\}, k \in \mathcal{K}_m$
While $\mathcal{N} \neq \emptyset$ **and** $(\sum_{l \in \{l_k^{y,m}\}} R^l > 0)$
  Find $(m^*, k^*) = \arg\min_{m,k} \ (\hat{W}_k - W_{k1}^m)$
  $l^* = s_{k^*}^{m^*}$
  Initialize $r^{l^*} = 0$
  While $\mathcal{N} \neq \emptyset$ **and** $r^{l^*} \leq R^{l^*}$
    Find $n^* = \arg\max_{n \in \mathcal{N}} e_n^{l^*}$
    $x_{n^*}^{l^*} = 1$
    $q_{k^*}^{m^*} = q_{k^*}^{m^*} - \min(r_{n^*}^{l^*}, q_{k^*}^{m^*})$
    $\mathcal{N} = \mathcal{N} - n^*$, $r^{l^*} = r^{l^*} + r_{n^*}^{l^*}$
  End
  $R^{l^*} = 0$
End
While $\mathcal{N} \neq \emptyset$ **and** $(\sum_{m=0}^{M} \sum_{k \in \mathcal{K}_m^o} q_k^m > 0)$
  Compute $u_{kn}^{d,m} = \alpha_k^{d,m} \beta_k q_k^m r_{kn}^{d,m}$, $m = 0, ..., M$, $k \in \mathcal{K}_m^o$
  Compute $u_{kn}^{f,m} = \alpha_k^{f,m} \beta_k (q_k^0 - Q_k^m) r_{kn}^{f,m}$, $m = 1, ..., M$, $k \in \mathcal{K}_m^o$
  Find $(y^*, m^*, k^*, n^*) = \arg\max_{y,m,k,n} u_{kn}^{y,m}$
  $x_{k^*,n^*}^{y^*,m^*} = 1$
  If $y^* = d$
    $q_{k^*}^{m^*} = q_{k^*}^{m^*} - \min(q_{k^*}^{m^*}, r_{k^*n^*}^{y^*,m^*})$
  Else
    $q_{k^*}^0 = q_{k^*}^0 - \min(q_{k^*}^0, r_{k^*n^*}^{y^*,m^*})$
  End If
  $\mathcal{N} = \mathcal{N} - n^*$
End

initializes the variables q_k^m based on Q_k^m and updates their values in each iteration, to take into account the effects of subchannel allocations in the next iterations. The reason for not using Q_k^m instead is to prevent any ambiguity, as the actual update of the queue sizes in the system should be done after the completion of the subchannel allocations, determination of the transmission rates over the links, and packet drops from the queues.

Note that the FDRA algorithm consists of two main blocks. Block 1 allocates subchan-

nels through lines 2 to 13 to meet the minimum rates specified, and block 2 allocates the subchannels through lines 14 to 25 to maximize the utility function. Since SUMR addresses delay-sensitive users by defining a minimum rate requirement for them whenever they have packets in their queues in the BS or relays and defines the utility function based on the delay-tolerant users, it thus uses block 1 for subchannel allocations to delay-sensitive users and block 2 for delay-tolerant users. On the other hand, JUMR allocates subchannels to all of the users based on block 2 and uses block 1 only when there are packets that have reached the defined deadlines on the links and (therefore) a minimum rate has been determined for them.

In all the policies, after allocating the resources and transmitting the data, if there is still any packet that has reached its deadline and not yet been transmitted (due to lack of resources), it is dropped from its queue. Then, based on the arrived, transmitted and dropped packets, the sizes and delays of the remaining packets are updated.

3.4.2 Enhanced CQDA Policies

In this subsection, we propose other variants for SUMR and JUMR policies, by making some modifications in the steps they take. As illustrated in the next section, these versions of the policies show improved performance compared to the original SUMR and JUMR policies. However, the original policies are still useful as they have better performance than the existing methods and need less computation or information than the enhanced policies.

a) **Enhanced SUMR (ESUMR).** This policy is similar to SUMR in steps 1 and 2; however, in step 3 it uses different prioritizing rule to provide the minimum rates.

Specifically, it replaces line 4 in the FDRA algorithm with the following:

$$l^* = \begin{cases} s_{k^*}^{m^*}, & \text{if } \hat{W}_{k^*} - W_{k^*1}^{m^*} = 0 \\ \arg\max_l \{\frac{\tilde{r}^l}{\bar{r}^l} R^l\}, & \text{otherwise} \end{cases} \tag{3.12}$$

where $\tilde{r}^l = \frac{1}{N}\sum_{n=1}^{N} r_n^l$ is the average of the achievable rates of the link l on the subchannels in the current time slot (and therefore gives a measure of the average current fading conditions that the link l has on different subchannels).

ESUMR aims at being more opportunistic. By using (3.12), it provides the minimum rates first for the links serving the packets that have reached their deadlines; after allocating subchannels to them, it gives priority based on the second line in (3.12), to utilize the opportunities of favorable channel states (through $\frac{\tilde{r}^l}{\bar{r}^l}$) while taking into account the queue and delay states (through R^l, which has been computed by (3.10) and is based on the packets' sizes and delays).

b) **Enhanced JUMR (EJUMR).** JUMR serves the delay-sensitive users mostly through maximizing the utility and giving higher priorities to them compared with delay-tolerant users. However, if the data arrival rates of the delay-tolerant users are large, their queues will get backlogged more frequently and their weights can overshadow the priority of the delay-sensitive users. EJUMR aims at preventing this, by taking into account the data arrival rates of all the traffic classes. In particular, it defines the β coefficients in step 1 using the following equation:

$$\beta_k = \frac{1}{\hat{W}_k} \max(1, \frac{\hat{\lambda}^b}{\lambda_k}), k \in \mathcal{K}^a \tag{3.13}$$

where λ_k is the average arrival rate in the queue of user k in the BS, and $\hat{\lambda}^b = \max_{k \in \mathcal{K}^b} \lambda_k$. According to (3.13), when the arrival rates of the delay-tolerant users are more than

that of a delay-sensitive user k, EJUMR increases the β_k in proportion to the ratio of $\hat{\lambda}^b$ and λ_k. This way, user k is able to maintain its priority through $\frac{1}{W_k}$.

3.4.3 Practical Considerations

In this subsection, we discuss the concerns that might arise about the assumptions and computational complexity of the proposed policies. First we note that the relays with buffering capability have been considered in other works [33, 47], and in practice, they are no different from nodes with store-and-forward capabilities commonly employed in wireless mesh or ad hoc networks. Also, the quasi full-duplex relaying can be realized in practice by using two antennas and two radios in the relay, a directional antenna for feeder links from the BS and an omnidirectional antenna for direct links to users [47]. Proper configuration and installation of these two independent radio frequency chains can minimize any potential interference between transmissions and receptions on subchannels over the two links at either sides of the relay, which are close in frequency. Furthermore, the proposed policies can also be easily extended for half-duplex relays; one simple extension would be to follow steps 1 and 2 of the proposed policies, to formulate the problem. Then in step 3, set $x_{kn}^{d,m} = 0, m = 1, ..., M, k \in \mathcal{K}_m$ for first half of the time slot and $x_{kn}^{f,m} = 0, m = 1, ..., M, k \in \mathcal{K}_m$ for second half, to respectively exclude the direct links from relays and feeder links to relays in FDRA algorithm.

As it can be seen through the proposed policies, the BS uses the information about the queue sizes, packet delays and channel states to formulate and solve the resource allocation problem. Queue sizes and packet delays in the BS can be easily obtained in practice from local information in the BS. Obviously the BS also has information about the queue sizes and packet delays in relay buffers, as it knows the transmission sizes from all the queues and over all the links in the previous time slot. On the other hand, for CSI, the users and

relays would need to send feedback to the BS about the quality of their links on different subchannels. In practice this is done by reporting the index of achievable modulation and coding schemes on the subchannels. Due to the limited number of these schemes, the overhead will be lower compared with reporting on a continuous range. Also, to further reduce the overhead, the BS can determine a threshold and ask the users and relays to report only channel states of the subchannels that have better quality than the specified threshold. We note that by considering perfect channel information, our policies can be used as a benchmark to which other policies can be compared. In the case of imperfect channel information, the performance of the proposed policies may be degraded; however, the relative improvement compared with the existing algorithms would hold, as perfect information is assumed in performance evaluations for existing algorithms as well.

Define $\Lambda = \max_{k \in \mathcal{K}} |\mathcal{M}_k| \leq M$ as the maximum number of serving nodes a user might have. The computations of steps 1 and 2, which are related to defining the parameters of IRAP formulation, and through lines 2-13 of the FDRA algorithm in step 3 are relatively insignificant and may be ignored. Thus the computational complexity of the proposed policies are mainly affected by the lines 14-25 of the FDRA algorithm and is of $O(K\Lambda N^2)$. This is a polynomial time complexity that makes the proposed polices suitable for practical implementations.

3.5 Performance Evaluation and Discussion

In this section, we investigate the performance of the policies described in the previous section. We consider a system with cell radius equal to 1000 m and with 6 relays located at a distance of 2/3 cell radius from the BS, with uniform angular distance from each other. BS, relay and user antenna heights are considered 15 m, 10 m and 1.5 m respectively and the path loss model is based on [61]. The noise power spectral density is assumed -174

dBm/Hz and BER requirements for delay-sensitive and delay-tolerant users are considered 10^{-4} and 10^{-6}, respectively. We consider a large subchannel bandwidth, 180 kHz, that is equal to that of resource blocks in the LTE standard, as it leads to a lower overhead for channel states reporting by receivers, less computations for resource allocations, and a lower overhead for announcing subchannel allocations decisions. Also for similar reasons, we consider a time slot duration equal to that of an LTE subframe, i.e., 1 ms. The carrier frequency is assumed 1900 MHz. For the transmit power at the BS or a relay on each subchannel, we down-scale the 43 dBm that is typical for a BS and maximum for a relay in LTE [66] by 100 times or 20 dB, which corresponds to the case of equal distribution of power on 100 LTE resource blocks or subchannels. Note that several works in the literature have considered equal maximum power for the BS and relays [23, 24, 66]. However, even in the cases that they do not have equal maximum power, our policies can be applied in two possible ways. Firstly, the BS and relays can allocate the same power to serve the users with QoS requirements and use the remaining power for BE services. Second, our proposed policies can be applied based on the relays' lowest power, such that subchannel allocations are made with underestimated achievable rates. Then each of the BS and relays can distribute their maximum power on the subchannels allocated to them, which will result in a better performance than the underestimated one.

For the links from the BS or relays to users, we consider a channel model with Rayleigh fading, and for the links between the BS and relays, a Rician channel model is used with κ factor equal to 6 dB [62]. Packet arrivals at the BS queues of delay-tolerant users are modeled as Poisson processes with the packet sizes equal to 1 kbits and various arrival rates that will be explained later. For the delay-sensitive users, we use the "MPEG Video" traffic model based on [64] such that the frame interarrival time is exponentially distributed with 40 time slots, number of packets in a frame is 4 and the packet interarrival time in

a frame is 2 ms. Packet sizes are exponentially distributed with 1 kbits, which result in a video data rate of 100 kbps, and their delay constraint is 150 ms.

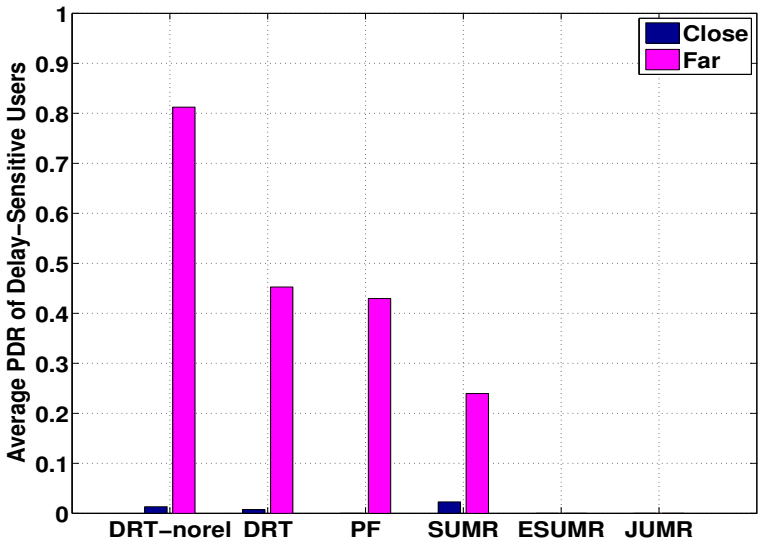
For comparison purposes, we also consider QoS provisioning in a system with no relay as well as a system with prompt relays (i.e., relays without buffer) in which, in the first half of a time slot, the BS transmits to relays and in the second half, relays forward the received data to the users. For providing QoS in the system with prompt relays, we adapt the dynamic resource allocation algorithm in [24] for non-cooperative transmissions, referred as DRT in the figures. In DRT algorithm, relay selection and subchannel allocation are performed through utility maximization, where the utility is defined based on the packet delays in the BS queues and the end-to-end relay channels' transmission rates, and minimum rate provisioning. For the case with no relay, referred to as "DRT-norel" in the figures, the algorithm of [24] is adapted such that all the users can receive data only from the BS. Also we have considered Proportional Fair (PF) scheduling in a system with buffering capability in relays. For this purpose, we adapt the algorithm in [67] such that the first half and second half of time slots are allocated to the BS or relays based on their needs to serve their queues, and subchannel assignment to users are performed based on the end-to-end PF metrics of the users.

To investigate the effectiveness of the proposed policies in providing delay guarantees for delay-sensitive users and average throughput for delay-tolerant users, first we consider some special scenarios with extreme limiting factors. Then we provide general results based on more typical scenarios. Note that except for the last scenario, the EJUMR policy is not illustrated in the figures, as it has the same performance as JUMR in these scenarios. This will become clearer as we explain the results.

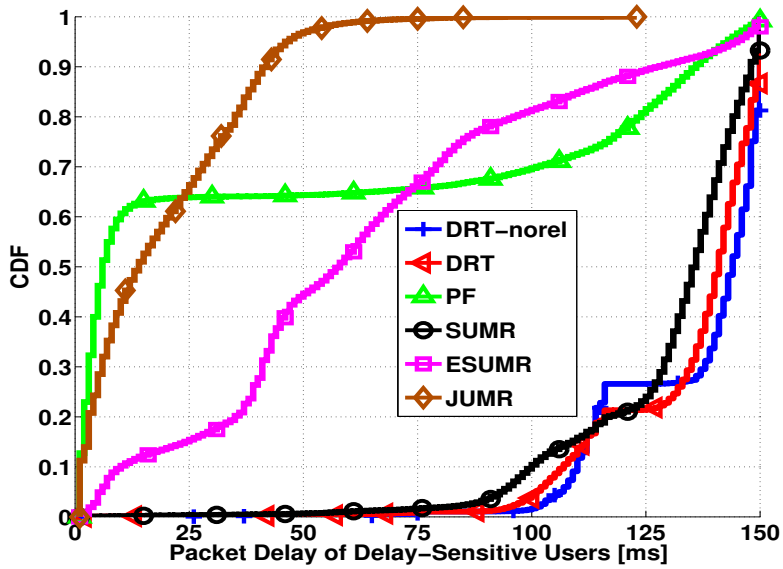
In order to see the delay guarantee performance for both the users with strong and weak wireless links, we consider a scenario with 3 subchannels and 12 users, all delay-sensitive;

6 of the users are close to the BS at 50m, and are connected directly to the BS; the other 6 are far from the BS at 1000m and located at an equal angular distance from each other, such that they have the same distance from the two nearest relays. Therefore, in addition to the BS, each of the 6 far users is also connected to two relays, which provide signals with higher average SNR than that from the BS. Fig. 3.3 shows the average Packet Drop Ratio (PDR) of the mentioned close and far users, as well as the CDF of the delay of the received packets (i.e., the dropped packets are not included). It is observed from Fig. 3.3(a) that the cases with no relay, prompt relays, as well as PF scheduling in a system with buffering relays are able to keep the PDR reasonably low or at zero for the close users, but suffer from high PDRs for far users. SUMR is able to reduce the PDR remarkably for the far users at the cost of a small increase in PDRs for close users. Overall, SUMR improves the PDR performance. ESUMR and JUMR have even better performance and result in zero PDR for all the users. Also, Fig. 3.3(b) shows that the packets that are successfully received by their destinations (i.e., they have not been dropped) experience lower delays by the proposed policies, especially with ESUMR and JUMR.

These results might look strange at the first glance, because when using two-hop transmissions, the packets might experience some queuing delay in the buffers of relays, and they might be expected to have higher delays and drop ratios compared with the prompt relays. However, what happens is the opposite, and the reason is as follows. Since the buffering capability in the relays makes it possible to take advantage of the channel variations opportunistically, such a system has a larger capacity. The proposed policies are able to exploit this well and provide high service rates for the queues so that the overall effect is the faster transfer of packets from the BS to users or equivalently, lower PDRs and delays. Specifically, for routing the packets of far delay-sensitive users, whenever they have not reached their deadlines, SUMR and ESUMR decide based on (3.8); therefore, they are



(a)



(b)

Figure 3.3: (a) Average PDR (b) CDF of the delay of the received packets; with 3 sub-channels, 6 close and 6 far users, all delay-sensitive.

able to transmit the packets through the nodes that have higher transmission rates and/or lower queuing delays. Also, in the case of routing through relays, using (3.6), they give most portion of the delay budgets to the link that has lower average rates. Based on these and using (3.10) they compute the minimum rate constraints of the links dynamically. All of the above make it possible to formulate IRAP in accordance with the potentials of the system with buffering relays. Using the FDRA algorithm then, they are able to utilize the system subchannels efficiently. However, since SUMR starts the provisioning of minimum rates based on line 4 in FDRA algorithm, it restricts the utilization of link diversity. On the other hand, ESUMR replaces line 4 with (3.12), which enables it to consider the channel conditions and minimum rate requirements together (except when there is an urgent situation for packet transmission), and function even more opportunistically. This way, it first transmits packets on the links with better conditions and/or more rate requirements. Therefore, it does not encounter the cases with many due packets that cannot be served altogether in a single time slot. As the above figures show, JUMR has the best performance among the proposed policies. This is achieved because in formulating the instantaneous problem, JUMR works mostly through utility maximization, and activates the minimum rate constraints only when necessary. As a consequence, routing and scheduling of the delay-sensitive packets is automatically done based on subchannel allocations through lines 14-25 in the FDRA algorithm. Since the utility function is based on MW policy, JUMR is able to perform the routing and subchannel allocations in a manner that takes advantage of the maximum capacity property of MW as much as possible, which translates into lower delays and PDRs over time.

Note that the improvement in the proposed algorithms is partly due to the fact that having buffers makes it possible to utilize the channel diversities better. Therefore, it is expected to have diminishing gains as the channel diversities increase. To show this, we

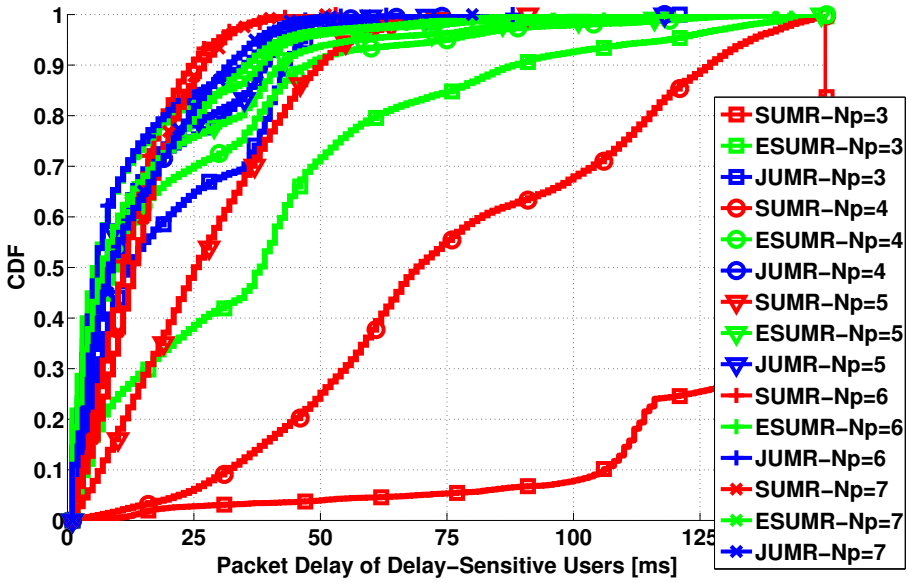
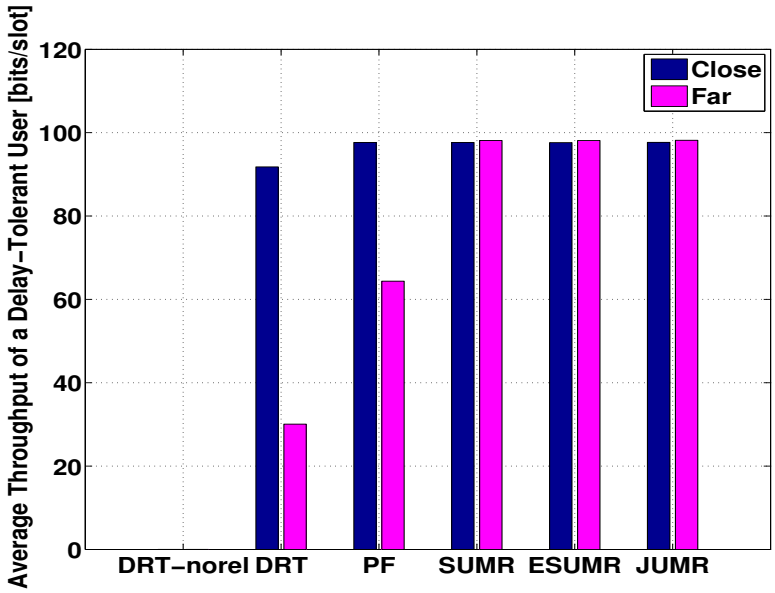


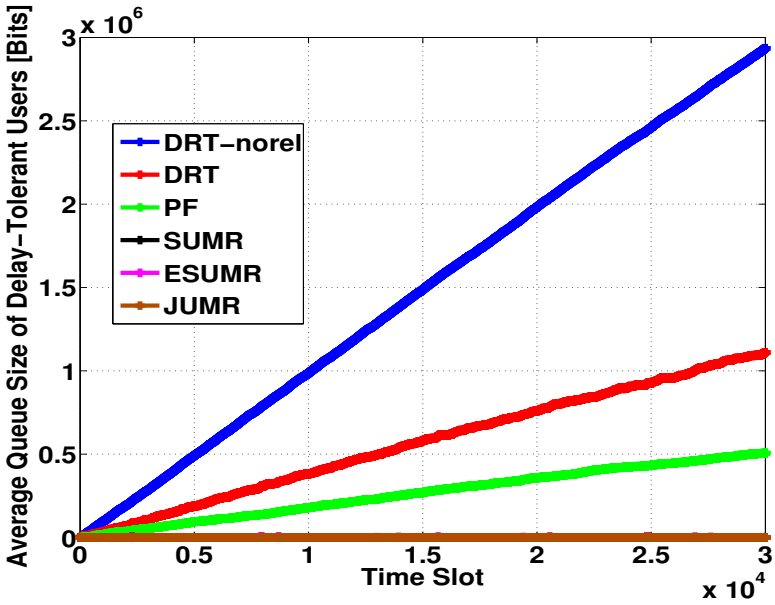
Figure 3.4: CDF of the delay of the received packets; with 3 allocatable subchannels from a pool of subchannels, 6 close and 6 far users, all delay-sensitive.

have conducted simulations for the scenario described above, with the difference that the 3 subchannels that can be allocated to users are selected from a pool of Np subchannels. Fig. 3.4 shows the delay CDF of the received packets with the proposed policies for different values of Np. The improvement in the delays, especially for SUMR, is largest when Np changes from 3 to 4. After that, while there is still reduction in the delays, this reduction becomes less as Np increases.

Next, we consider a similar scenario with the difference that there are 4 subchannels and also half of the users in each group, close and far, have delay-tolerant traffic with the average rates equal to 100 kbps. Fig. 3.5 illustrates the average throughput in each time slot and the average queue size over time, for delay-tolerant users. It is observed that the proposed policies are able to keep the queues of the delay-tolerant users stable and, therefore, provide them with the average throughput equal to the average data arrival rates in their queues in the BS. With DRT-norel and DRT, the average queue sizes increase



(a)

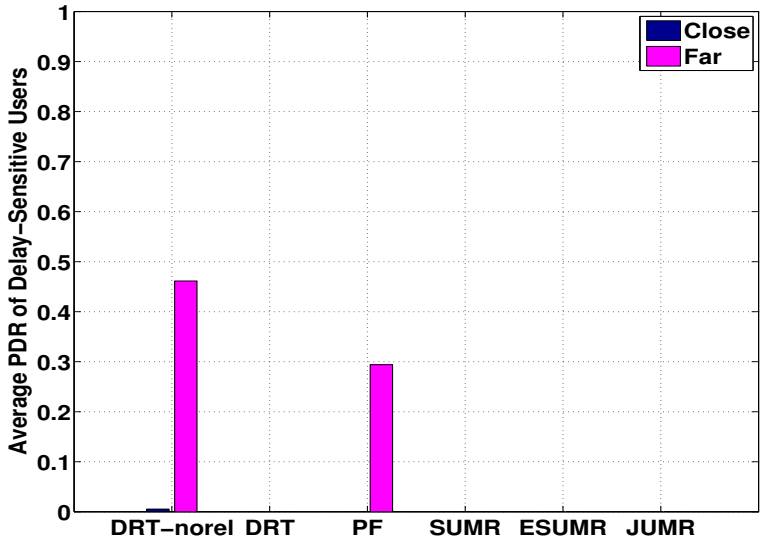


(b)

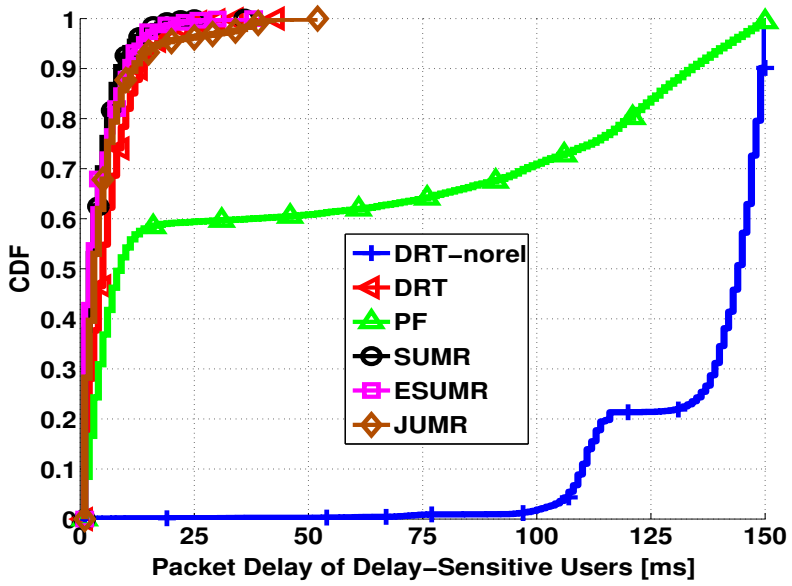
Figure 3.5: (a) Average throughput (b) average queue size of the delay-tolerant users; with 4 subchannels, 6 close and 6 far users, half of the users in each group are delay-sensitive and the other half are delay-tolerant.

steadily, which leads to instability and therefore, lower throughput for the close and far users. With PF, while the queues of close users remain stable and they get an average throughput equal to their data arrival rates at the BS, the queues of far users suffer from instability resulting in lower throughput for these users. In order to get a complete image of the results, the PDR and the delay CDF of the received packets for delay-sensitive users are also shown in Fig. 3.6. All the policies, except for DRT-norel (which has the smallest capacity) and PF (which is not QoS-aware), are able to keep the PDR zero in this scenario. DRT achieves this at the cost of queue instability for delay-tolerant users (especially the far users, as shown in Fig. 3.5(a)), but the proposed policies utilize the buffer-aided relaying system's capacity well to satisfy both the delay guarantee and queue stability objectives, for both close and far users. In particular, considering the product of delay-tolerant users' queue sizes and transmission rates in the utility function (3.4a) enables CQDA policies to serve the queues of delay-tolerant users in the BS and relays when their queue sizes are large and/or their links have good channel conditions, which contribute to their stability. The delay guarantees of delay-sensitive packets are also obtained due to the same reasons as in Fig. 3.3.

In the following we consider some general scenarios, where users are randomly distributed in the cell area with uniform distribution. The number of users is set to $K = 22$ (as before, half of them are delay-sensitive and half, delay-tolerant) and the number of subchannels is set to $N = 10$. We have conducted simulations for 100 different distribution of the user locations (resulting in totally 1100 different individual locations for a delay-sensitive user and the same number for a delay-tolerant one), each over 10000 time slots. The simulation results obtained in this scenario showed that except for DRT-norel and PF (which resulted in user PDRs in the ranges [0, 0.8] and [0, 1], respectively), DRT and our proposed policies were able to keep PDR equal to 0. We have not shown the figure for



(a)

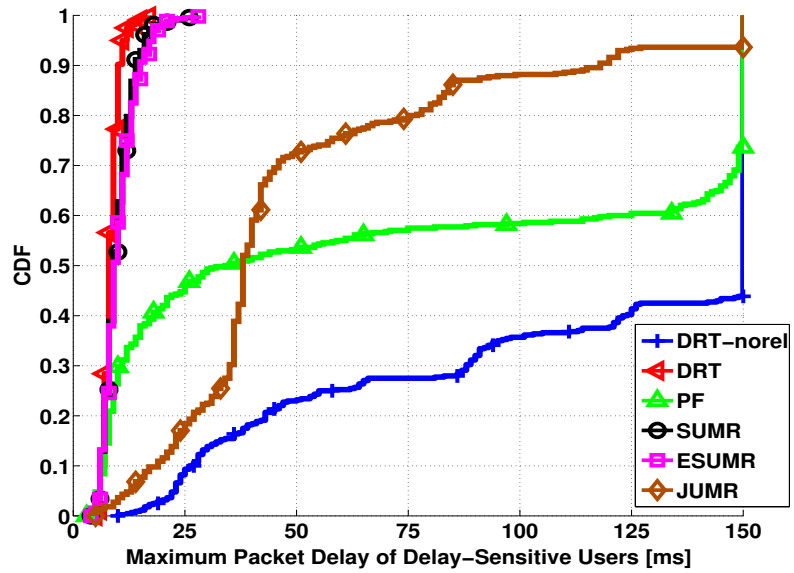


(b)

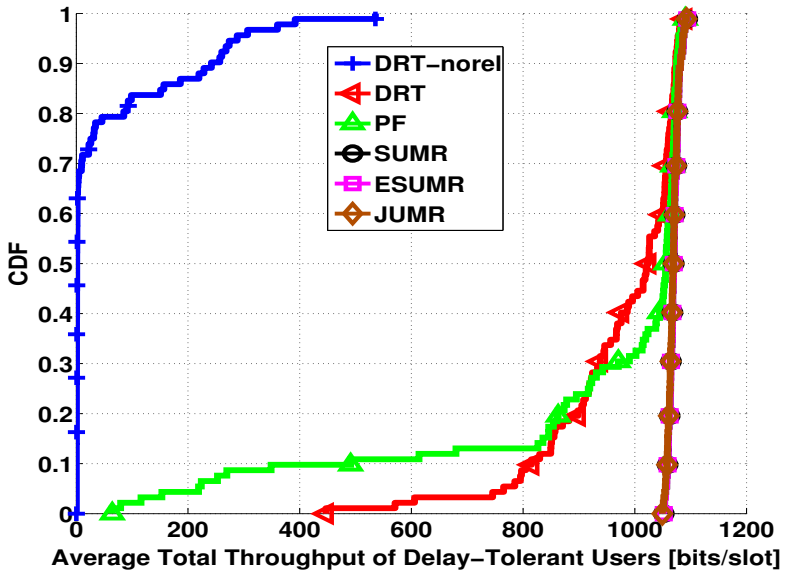
Figure 3.6: (a) Average PDR (b) CDF of the delay of the received packets for the delay-sensitive users; with 4 subchannels, 6 close and 6 far users, half of the users in each group are delay-sensitive and the other half are delay-tolerant.

this, as it did not give much more information than what has been mentioned. Instead, Fig. 3.7 displays the CDF of the maximum delay of the received packets for delay-sensitive users as well as the CDF of the average total throughput of delay-tolerant users in each time slot. It is observed that the proposed policies are able to meet the deadline of all the delay-sensitive packets and at the same time provide queue stability and, therefore, average total throughput rate equal 1100 bits per time slot, which is equal to the sum of average data arrival rates at the queues of 11 delay-tolerant users in the BS. On the contrary, DRT-norel and DRT lead to instability and, therefore, less throughput than what is injected into network, as they favor the delay-sensitive users and try to meet the deadlines of their packets. PF is able to provide stability in some realizations of the users' locations in the cell area, but in some other realizations it leads to instability and throughputs far below the data arrival rates. It is firstly due to the fact that PF does not use the system resources efficiently and secondly because it tries to provide fair service to all the users without taking into account the different requirements of them.

Fig. 3.7(a) also shows that the packets of delay-sensitive users experience more delay with JUMR than with SUMR, ESUMR or even DRT. The reason is that the IRAP formulations of SUMR, ESUMR and DRT consider the serving of delay-sensitive users by defining minimum rate constraints; this causes the FDRA algorithm (through lines 2-13) to allocate the subchannels in every time slot, first to delay-sensitive users and then, if any left, to delay-tolerant users. On the other hand, with the JUMR problem formulation, the delay-sensitive users' service are mostly considered based on utility maximization; therefore, when the utility of delay-tolerant users get higher than that of delay-sensitive users in some time slots (because of possibly larger queue sizes or channel gains), subchannels are allocated first to them (through lines 14-25 in FDRA). This postpones the service of delay-sensitive packets and results in larger delays for them.



(a)

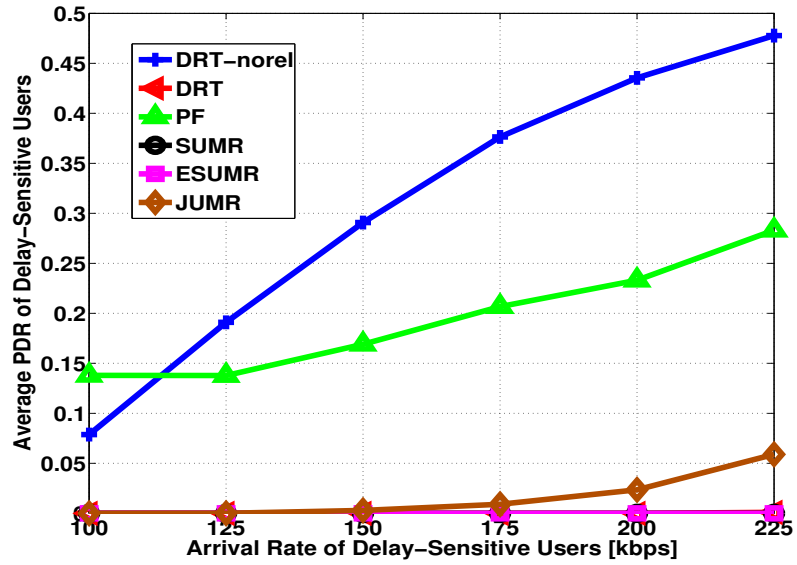


(b)

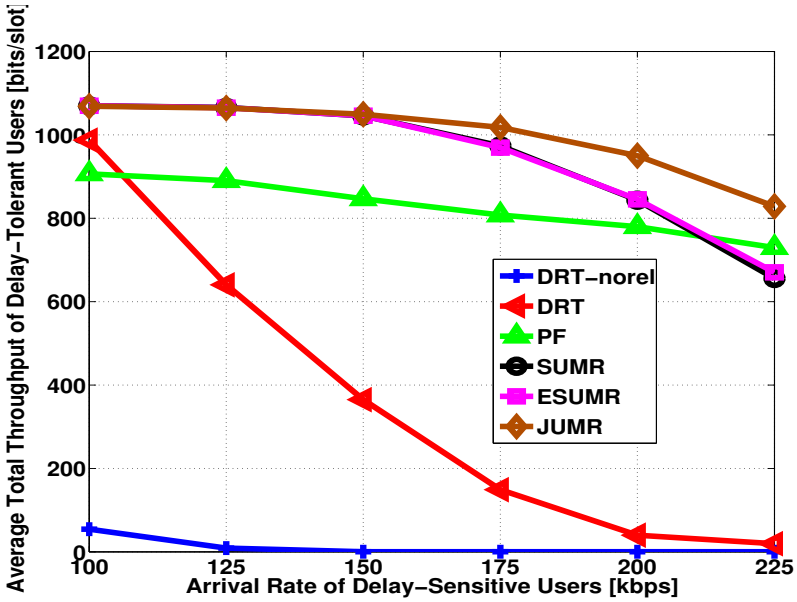
Figure 3.7: (a) Maximum packet delay of delay-sensitive users (b) average total throughput of delay-tolerant users; $N = 10$, $K = 22$, random distribution of users in the cell area.

Note that PDR=0 for some policies in the results above only shows the relative effectiveness of these policies compared with other policies and is not an indicator of optimality. As the system load increases, the PDR might become nonzero. This is clarified through the following evaluations. In a scenario same as above, we fix the data arrival rate of delay-tolerant users at 100 kbps and investigate the effect of increasing the traffic load of delay-sensitive users. Fig. 3.8 shows that as the arrival rate of delay-sensitive packets increases, DRT-norel leads to higher PDR and almost stops providing data throughput for delay-tolerant users at the video rates larger than 125 kbps. With PF, due to QoS unawareness and inefficiency, PDR of delay-sensitive users increases and throughput of delay-tolerant users decreases. However, due to buffering capability in relays and larger system capacity, the slop of changes is low. On the other hand, with the increase in the data arrival rate of delay-sensitive users, DRT, SUMR and ESUMR define larger values for minimum rate requirements, and use larger amount of resources (subchannels over time) to meet the packet deadlines and keep the PDR equal to zero. The consequence is that DRT (which is unstable even at the video arrival rates of 100 kbps), reduces the service rates of delay-tolerant users quickly and provides very low throughput to them at the video rates of 225 kbps. However, SUMR and ESUMR are able to keep the queues of delay-tolerant users stable and support their data arrival rate (i.e., provide the throughput equal to 100 kbps) as long as the data arrival rate of delay-sensitive users is not larger than 150 kbps. JUMR responses differently, due to its different approach. Although it results in the instability of the delay-tolerant users' queues as the video rate gets larger than 150 kbps, the reduction in the throughput of delay-tolerant users is less, compared with other policies. Instead, JUMR penalizes the delay-sensitive users by dropping their packets.

Now we consider a similar scenario, with the difference that the data arrival rates of the delay-sensitive users are fixed at 100 kbps and the effect of the increase in the traffic



(a)

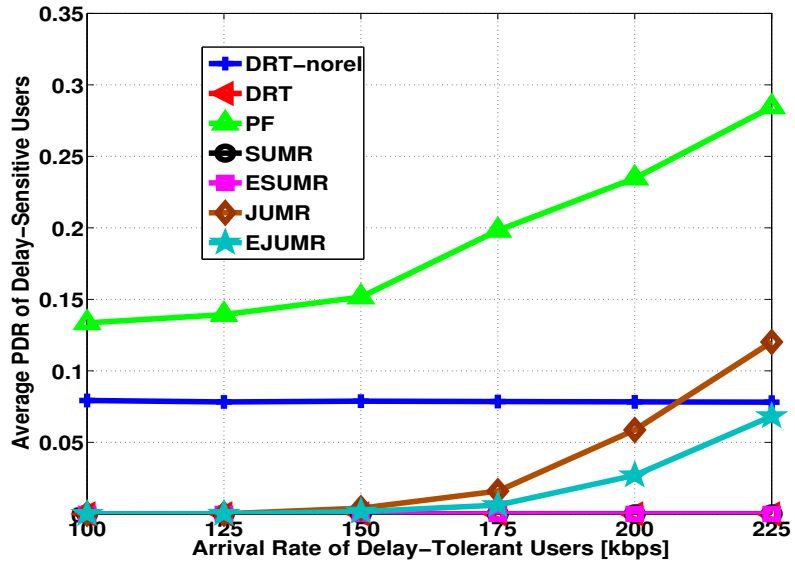


(b)

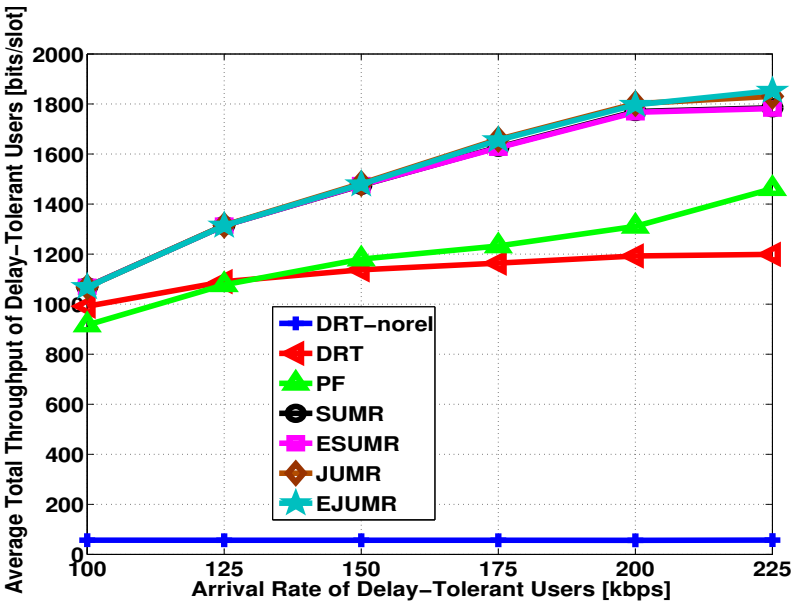
Figure 3.8: Effect of data arrival rate of delay-sensitive users on (a) average PDR of delay-sensitive users (b) average total throughput of delay-tolerant users.

load of delay-tolerant users is investigated. In this case, it is also possible to illustrate the performance of EJUMR policy more clearly. Fig. 3.9 shows that as the data arrival rate of delay-tolerant users increases, PF results in higher PDRs for delay-sensitive users. This is due to the fact that it takes into account the queue sizes in allocating the resources and leads to more service to delay-tolerant users (due to the increase in their data arrival rates and queue sizes), the result of which is increase in these users' throughputs. DRT-norel shows no change in the PDR of delay-sensitive users and in the throughput of delay-tolerant users. This happens because the arrival rates of video traffic are fixed and DRT-norel uses a fixed amount of resources for them. Therefore, a fixed amount is left for delay-tolerant users, which is not enough to stabilize the system even under a data arrival rate of 100 kbps. As a result, it provides a fixed throughput for delay-tolerant users, no matter how much higher the arrival rate of their traffic is. Due to similar reasons, DRT, SUMR and ESUMR consume a fixed (but different) amount of resources for delay-sensitive users and thanks to larger capacities achieved from using the relays, they are able to meet the deadline of all the delay-sensitive packets. However, while DRT is unstable even at the arrival rate of 100 kbps, SUMR and ESUMR are able to keep the system stable up to the arrival rate of 150 kbps by matching the throughput of delay-tolerant users to the arrival rates of their traffic.

With JUMR and EJUMR, the system also remains stable and keeps the PDR equal to zero, up to a data arrival rate of 150 kbps; but as the arrival rate increases further, they provide higher throughput than SUMR and ESUMR and instead, drop some of the delay-sensitive packets as well. As mentioned in Section 3.3, EJUMR takes into account the data arrival rates of the users when defining their weights in the utility function; therefore, it is able to maintain the priority of delay-sensitive users better than JUMR, which results in lower PDRs. Based on this and the previous observations, it is inferred that EJUMR



(a)



(b)

Figure 3.9: Effect of data arrival rate of delay-tolerant users on (a) average PDR of delay-sensitive users (b) average total throughput of delay-tolerant users.

(which has the same performance as JUMR in the previous scenarios) is better able to utilize the system resources to provide system stability and throughput for delay-tolerant users, while providing delay guarantees for delay-sensitive users. When the system starts to get unstable due to traffic overload, EJUMR degrades the service of both delay-sensitive users and delay-tolerant users; however, to return the system to a stable state, it will need a smaller reduction in the service quality of users, compared with the other policies, since it has a larger capacity.

3.6 Summary

In this chapter, we have provided a novel framework for formulating and solving QoS-aware resource allocation in OFDMA networks enhanced with buffering relays, with the objective to guarantee throughput and stringent packet delays, respectively for delay-tolerant and delay-sensitive users. We have shown that due to the multihop structure of the network and heterogeneous requirements of user services, the frequency domain resource allocation problem is not clear. We have presented the challenges and proposed CQDA policies to address them using different approaches, namely SUMR and JUMR. SUMR aims at serving the delay-sensitive users by defining minimum rate constraints and the delay-tolerant users by defining utilities for them, whereas JUMR defines utilities for all the users and activates minimum rate constraints only when getting close to packet deadlines of delay-sensitive users. We have also proposed enhanced versions of these policies, as ESUMR and EJUMR. Using extensive simulations, we have evaluated the performances of the proposed policies in different scenarios. The results show significant improvements in terms of packet delay guarantees and throughput provisioning, compared with the systems without relays, systems assisted by relays without buffering capability, as well as systems with buffering relays but QoS-unaware.

Chapter 4

Dynamic Distributed Resource Allocation Framework

4.1 Introduction

Resource allocation in relay-assisted cellular networks might be implemented in a centralized or distributed way. In the case of centralized implementation, the scheduler requires the information about users' queue states in the BS and the channel conditions of all the system links. Obviously, it has the information about the queue states in the relays, as it knows the history of the transmissions from all the queues in the system. On the other hand, distributed resource allocation is of more interest in cellular networks. This is firstly due to the reduced computational burden on the BS. Secondly, it is because of the fact that the signaling overhead of CSI is lower in the distributed scenarios.

In the previous chapter, due to the heterogeneous service requirements of users and strict deadlines of delay-sensitive packets, we were concerned more about QoS provisioning rather than implementation complexity and therefore, we proposed centralized policies. In this chapter, we consider only delay-tolerant services and aim at designing a framework for distributed resource allocation. We formulate power and subchannel allocation as an optimization problem and by introducing some concepts, we show its similarity to a multicell OFDMA scenario with smaller cells. Moreover, to make the problem tractable, we transform it into a convex optimization problem and using dual decomposition, we propose

a dynamic distributed iterative algorithm, called DDRA, where the BS and relays solve their own problem based on their links' QCSI and some global variables exchanged among them. DDRA provides a novel framework for exploiting the system's power and subchannel resources in an adaptive way over time, with lower overhead of the CSI feedback and lower computational complexity at the BS compared with optimal centralized scheduling which requires global CSI at the BS.

The rest of this chapter is organized as follows. In Section 4.2, we outline the system model for a relay-assisted OFDMA network. In Section 4.3, we formulate the resource allocation algorithm design as an optimization problem and solve it by dual decomposition, where distributed closed-form solutions for power and subcarrier allocation are derived. Simulation results are studied in Section 4.4, and finally, the conclusion is presented in Section 4.5.

4.2 System Model

We consider a single cell time slotted OFDMA system in the downlink, with K users and M relays. Users are randomly distributed in the cell area, K_1 of them being served directly by the BS while others receive data through one of the relays. As it is shown in Fig. 4.1, we assume that each user has been assigned to either the BS or any of the relays based on a criteria such as average SNR, distance from the BS and relays, etc.

BS and relays are equipped with buffers, where the BS has one for each user but relays have one for each of only the users connected to them. Users' packets arrive at the BS buffer according to their traffic model and are queued until transmission to the directly connected users or to the relays serving other users. Relays do not need to transmit the received packets immediately in the next time slot and it is possible to keep them in the buffers and serve them based on the scheduling policy. This gives flexibility to the scheduler

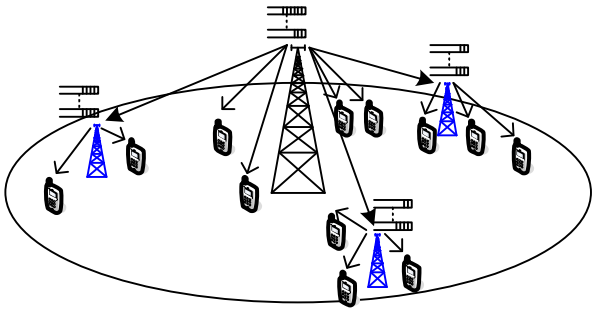


Figure 4.1: System model

to utilize the resources more opportunistically by postponing the transmission until the user gets higher priority or better channel. We use Q_k^B, $k = 1, \ldots, K$ to denote the queue size of user k in the BS, and $Q_k^{R(k)}$, $k = K_1 + 1, \ldots, K$ to denote the queue size of user k in its serving relay, $R(k)$.

We assume that the transmission bandwidth is divided into N subchannels where each subchannel can be used exclusively by the BS or relays in one of the groups of the links, i.e., BS-to-users, BS-to-relays and relays-to-users. Any relay has the ability to transmit on some subchannels and at the same time receive data from the BS on other subchannels. The channel conditions of all the links are assumed time variant and frequency selective, but constant during one time slot and over one subchannel. We use e_{kn}^B to indicate the channel gain-to-noise ratio of the link between the BS and user k on subchannel n. Similarly, $e_{kn}^{R(k)}$ and $e_n^{BR(k)}$ denote the channel gain-to-noise ratio on subchannel n for the link between $R(k)$ and user k and the link between the BS and $R(k)$, respectively. Assuming that M-ary QAM modulation is used for transmission, the achievable transmission rates can be computed based on the following [63]:

$$r_{kn}^B = x_{kn}^B \log_2 \left(1 + \frac{p_{kn}^B e_{kn}^B}{\Gamma_k}\right), k = 1, \ldots, K_1, \tag{4.1}$$

where, without loss of generality, the bandwidth of a subchannel has been assumed equal

to 1. r_{kn}^{B} is the achievable transmission rate between the BS and user k on subchannel n. x_{kn}^{B} denotes subchannel allocation indicator, which is one if subchannel n is used by the BS to transmit data to user k, $k = 1...K_1$, and zero otherwise. p_{kn}^{B} is the power allocated by the BS to user k on subchannel n. Γ_k is the SNR gap due to the limited number of coding and modulation schemes and is related to bit error rate of user k (BER_k), through equation $\Gamma_k = -\frac{\ln(5BER_k)}{1.5}$. In a similar way we can define $x_{kn}^{R(k)}$,$p_{kn}^{R(k)}$ and $r_{kn}^{R(k)}$ for the links from the relays to users and $x_{kn}^{BR(k)}$, $p_{kn}^{BR(k)}$, and $r_{kn}^{BR(k)}$ for the links from the BS to relays. In each time slot, a resource allocation algorithm aims at efficient use of the system resources, i.e., power and subchannles, while considering the QoS for the users in terms of BER and queue stability. This is discussed in detail, in the next section.

4.3 Cross Layer Scheduling and Resource Allocation

In this section, we formulate the cross layer scheduling and resource allocation problem and then, using some definitions and modifications, we propose a new perspective with simplified convex optimization problem.

4.3.1 Problem Formulation

In each time slot, the resource allocation policy considers the optimization problem in (4.2) to decide about the allocation of system power and subchannels. In this problem, w_k^B, $w_k^{BR(k)}$, and $w_k^{R(k)}$ are the weights of the users over the links from the BS to users, the links from the BS to relays and the links from the relays to users, respectively. $C1$ states the total power constraint for the BS and the relays, where $P_t \geq 0$ is the total available power in the system. $C2$ indicates that each subchannel can be allocated to only one link. $C3$ and $C4$ define the feasible values for suchannel allocation indicators and the powers used on the subchannels, respectively.

$$P: \max_{\boldsymbol{p},\boldsymbol{x}} \quad \sum_{k=1}^{K_1}\sum_{n=1}^{N} w_k^B r_{kn}^B + \sum_{k=K_1+1}^{K}\sum_{n=1}^{N} w_k^{BR(k)} r_{kn}^{BR(k)} + \sum_{k=K_1+1}^{K}\sum_{n=1}^{N} w_k^{R(k)} r_{kn}^{R(k)}, \tag{4.2a}$$

$$\text{s.t. } C1: \sum_{n=1}^{N}\Big(\sum_{k=1}^{K_1} p_{kn}^B + \sum_{k=K_1+1}^{K} (p_{kn}^{BR(k)} + p_{kn}^{R(k)})\Big) \le P_t, \tag{4.2b}$$

$$C2: \sum_{k=1}^{K_1} x_{kn}^B + \sum_{k=K_1+1}^{K} (x_{kn}^{BR(k)} + x_{kn}^{R(k)}) \le 1, \forall n, \tag{4.2c}$$

$$C3: x_{kn}^B, x_{kn}^{BR(k)}, x_{kn}^{R(k)} \in \{0,1\}, \forall k, n, \tag{4.2d}$$

$$C4: p_{kn}^B, p_{kn}^{BR(k)}, p_{kn}^{R(k)} \ge 0, \forall k, n, \tag{4.2e}$$

The problem (4.2) is a mixed integer nonlinear programming problem which needs an exhaustive search to find the optimal solution. Note that it is feasible because the system total power is non-negative. In order to make the problem tractable, we relax the subchannel assignment variables $x_{kn}^B, x_{kn}^{BR(k)}, x_{kn}^{R(k)}$ to be real value between zero and one, instead of a Boolean, i.e., $0 \le x_{kn}^B, x_{kn}^{BR(k)}, x_{kn}^{R(k)} \le 1$, which is known as time or tone sharing [68]. Furthermore, we consider the buffers of users $k = K_1, \ldots, K$ in their relays, as virtual users that are directly connected to the BS. In other words, we interpret the links between the BS and relays as the direct links between the BS and some virtual users. As shown in Fig. 4.2, this perspective helps us to divide the serving area (single cell) into smaller areas (multi cells) served by $M+1$ nodes, where node 0 is the BS with K users and has the complicated Radio Resource Management (RRM) capability and acts as a central controller while nodes $m, m = 1, \ldots, M$, are the relays with their own users, totally $K - K_1$ users, and act as antennas distributed in the serving area and connected wirelessly to the controller. We denote the set of users of node m with $\mathcal{Q}_m$; in particular $\mathcal{Q}_0 = 1...K$. Each

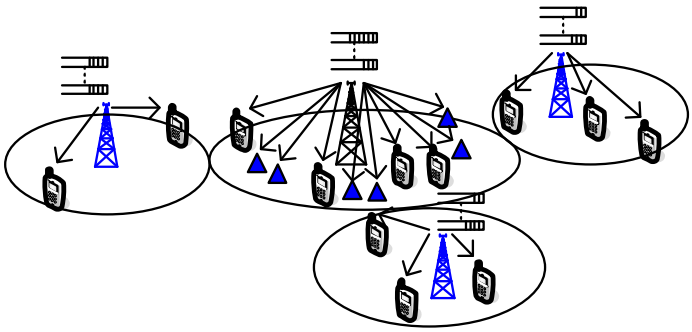


Figure 4.2: Similarity of the model to multicell network

node has the buffers of its own users and transmits data independently; however, in the beginning of each slot they all communicate with a central controller in node 0 to decide about their shares of power and subchannels and prevent interference on other nodes.

We use the following notations for each user:

$$e_{kn}^{m} = \begin{cases} e_{kn}^{B}, \quad m = 0, k = 1, \ldots, K_1 \\ e_{n}^{BR(k)}, m = 0, k = K_1 + 1, \ldots, K \\ e_{kn}^{R(k)}, m = 1, \ldots, M, k \in \mathcal{Q}_m \end{cases}$$

x_{kn}^{m} and p_{kn}^{m} can be defined in a similar way. We define $\mathcal{D} = \{(\boldsymbol{p}, \boldsymbol{x}) | 0 \leq p_{kn}^{m} \leq P_t, x_{kn}^{m} \in [0, 1]\}$ as the domain of the problem. Due to tone sharing, SNR is equal to $\frac{p_{kn}^{m} e_{kn}^{m}}{x_{kn}^{m}}$; this SNR is because of viewing p_{kn}^{m} as the energy per time slot that node m uses for user k on subchannel n [68]. As a result, the rates are computed by $r_{kn}^{m} = x_{kn}^{m} \log_2(1 + \frac{p_{kn}^{m} e_{kn}^{m}}{x_{kn}^{m} \Gamma_k})$.

Assuming that the system is stabilizable, using MW [54, 55], the queue stability can be provided by defining the weights of users as follows:

$$w_{k}^{m} = \begin{cases} Q_{k}^{B}, \quad m = 0, k = 1, \ldots, K_1 \\ Q_{k}^{B} - Q_{k}^{R(k)}, m = 0, k = K_1 + 1, \ldots, K \\ Q_{k}^{R(k)}, m = 1, \ldots, M, k \in \mathcal{Q}_m \end{cases} \tag{4.3}$$

Using the framework mentioned above, resource allocation problem is represented as:

$$\max_{\boldsymbol{p},\boldsymbol{x}\ \in\mathcal{D}} \sum_{m=0}^{M}\sum_{k\in\mathcal{Q}_m}\sum_{n=1}^{N} w_k^m x_{kn}^m \log_2(1+\frac{p_{kn}^m e_{kn}^m}{x_{kn}^m \Gamma_k}), \tag{4.4a}$$

$$\text{s.t.} \quad C1: \sum_{m=0}^{M}\sum_{k\in\mathcal{Q}_m}\sum_{n=1}^{N} p_{kn}^m \le P_t, \tag{4.4b}$$

$$C2: \sum_{m=0}^{M}\sum_{k\in\mathcal{Q}_m} x_{kn}^m \le 1, \forall n \tag{4.4c}$$

Note that the ordinary OFDMA networks can be considered as a special case of this formulation where M=0; in that case, the virtual users are the real users directly connected to the BS.

Problem 4.4 is convex and the strong duality holds [68, 69] (This can be verified by defining $\tilde{p}_{kn}^m = \frac{p_{kn}^m}{x_{kn}^m}$ and substituting in the objective and constraints). Therefore, using dual decomposition, an iterative algorithm can be designed to solve the problem.

4.3.2 Dual Problem Formulation

In this subsection, we formulate the dual problem for the resource allocation optimization problem. For this, we first obtain the Lagrangian function of primal problem. After rearranging the terms, the Lagrangian can be written as:

$$\begin{aligned}
\mathcal{L}(\boldsymbol{p},\boldsymbol{x},\mu,\boldsymbol{\delta}) &= \sum_{m=0}^{M}\sum_{k\in\mathcal{Q}_m}\sum_{n=1}^{N} w_k^m x_{kn}^m \log_2(1+\frac{p_{kn}^m e_{kn}^m}{x_{kn}^m \Gamma_k}) \\
&- \sum_{m=0}^{M}\sum_{k\in\mathcal{Q}_m}\sum_{n=1}^{N} \mu p_{kn}^m \\
&- \sum_{m=0}^{M}\sum_{k\in\mathcal{Q}_m}\sum_{n=1}^{N} \delta_n x_{kn}^m \\
&+ \mu P_t + \sum_{n=1}^{N}\delta_n
\end{aligned} \tag{4.5}$$

where $\mu \geq 0$ is the Lagrangian multiplier associated with the total power constraint, and $\boldsymbol{\delta} \geq 0$ is the Lagrangian multiplier vector for the subchannel allocation constraints. The dual problem is given by:

$$\min_{\boldsymbol{\mu},\boldsymbol{\delta}\geq 0} \max_{\boldsymbol{p},\boldsymbol{x}\in\mathcal{D}} \mathcal{L}(\boldsymbol{p},\boldsymbol{x},\mu,\boldsymbol{\delta}) \tag{4.6}$$

Similar to the method in [68], the dual problem can be solved by a centralized iterative algorithm in the BS. In this case, since the BS has the information of the previous transmissions, it would have the Queue State Information (QSI) of all the relays and therefore, there is no need for QSI signaling. On the other hand, it is needed to have all the users send feedback to their serving nodes about the CSI of their links. Then, the relays should inform the BS about the CSI of their local links, i.e., the links between themselves and their users. Moreover, they should send feedback to the BS about the CSI of the links between the BS and themselves.

Alternatively, using dual decomposition and concept of pricing, we propose an iterative distributed algorithm where in each iteration, the BS and relays solve their own problem based on the global variables and the weights and channel conditions of their local links. In this case, the relays should notify the BS about their queue sizes, to be used in the weights of the links between the BS and relays. For that, it suffices to report their modified queue sizes since the previous report, which leads to QSI signaling of at most in the order of $\mathrm{O}(\min(K-K_1, N))$. This is because there are totally $K-K_1$ queues in the relays and in each time slot, at most N different queues can be served by the system subchannels. On the other hand, CSI reporting is similar to the case of centralized resource allocation except that the relays do not need to inform the BS about the CSI of their local links; this reduces the CSI signaling overhead by the order of $\mathrm{O}((K-K_1)N)$.

In the following subsection, we study the distributed scheme and solve the dual problem in (4.6) by decomposing it into two parts: the first part is the local subproblem to be solved

by each of the serving nodes, BS and relays, and the second part is the main dual problem to be solved by the BS.

4.3.3 Dynamic Distributed Resource Allocation

By dual decomposition, the dual problem is decomposed into a main global problem and $M+1$ local subproblems, which can be solved iteratively. In each iteration, using the dual variables which are global for all the nodes, the BS and relays solve their local subproblem based on their QCSI. Then relays report their results to the BS and the BS updates the dual variables and broadcasts them to relays. In this way, dual variables act as prices that the BS adjusts to control the demands. The local subproblem in each node is given by:

$$\max_{\boldsymbol{p},\boldsymbol{x}\in\mathcal{D}} \mathcal{L}_m(\boldsymbol{p},\boldsymbol{x},\mu,\boldsymbol{\delta}),$$

where

$$\begin{aligned}\mathcal{L}_m(\boldsymbol{p},\boldsymbol{x},\mu,\boldsymbol{\delta}) &= \sum_{k\in\mathcal{Q}_m}\sum_{n=1}^{N} w_k^m x_{kn}^m \log_2(1+\frac{p_{kn}^m e_{kn}^m}{x_{kn}^m \Gamma_k}) \\ &- \sum_{k\in\mathcal{Q}_m}\sum_{n=1}^{N} \mu p_{kn}^m \\ &- \sum_{k\in\mathcal{Q}_m}\sum_{n=1}^{N} \delta_n x_{kn}^m \end{aligned} \tag{4.7}$$

where the Lagrange multipliers μ and $\boldsymbol{\delta}$ are provided by the BS. Using the Karush-Kuhn-Tucker conditions we obtain:

$$\frac{\partial \mathcal{L}_m}{\partial p_{kn}^m} = \frac{w_k^m x_{kn}^m e_{kn}^m}{\ln 2(x_{kn}^m \Gamma_k + p_{kn}^m e_{kn}^m)} - \mu = 0 \tag{4.8}$$

As a result, power allocation for subchannel n is obtained by:

$$
\begin{aligned}
& p_{kn}^{m\,*}(\boldsymbol{x}, \mu, \boldsymbol{\delta}) = x_{kn}^{m} \tilde{p}_{kn}^{m}(\mu), \\
& \tilde{p}_{kn}^{m}(\mu) = \min\Big(P_t, \Big(\frac{w_k^m}{\mu \ln 2} + \frac{\ln(5BER_k)}{1.5\ e_{kn}^m}\Big)^+\Big)
\end{aligned}
\tag{4.9}
$$

where $(a)^+ = \max(a, 0)$. After substituting $p_{kn}^{m\,*}$ into (4.7), we have:

$$
\begin{aligned}
\mathcal{L}_m \quad & (\boldsymbol{x}, \mu, \boldsymbol{\delta}) = \sum_{k \in \mathcal{Q}_m} \sum_{n=1}^{N} x_{kn}^m V_{kn}^m, \\
V_{kn}^m \quad & = w_k^m \log_2(1 + \frac{\tilde{p}_{kn}^m e_{kn}^m}{\Gamma_k}) - \Big(\mu \tilde{p}_{kn}^m + \delta_n\Big)
\end{aligned}
\tag{4.10}
$$

Define $V_n^{m*} = \max\limits_{k \in \mathcal{Q}_m} \{V_{kn}^m\}$; then, (4.10) is maximized if subchannel assignment variables are computed as follows:

$$
x_{kn}^{m\,*}(\mu, \boldsymbol{\delta}) = \begin{cases} 1, & V_{kn}^m = (V_n^{m*})^+, \\ 0, & V_{kn}^m < (V_n^{m*})^+, \end{cases}
\tag{4.11}
$$

In some time slots, more than one users might have $V_{kn}^m = (V_n^{m*})^+$. This happens mostly for the virtual users of the BS that represent the links belonging to the same group, i.e., between the BS and a particular relay, as these links have the same channel condition over a subchannel. In such cases, the tie can be broken arbitrarily. According to (4.3), (4.9) and (4.11), queue sizes of users, their channel conditions and required BER affect their share of power and subchennals.

4.3.4 Solution of Main Dual Problem at the BS

Using the information about the power and channel allocation variables reported by relays and based on the subgradient method [70], the BS updates the dual variables through the

following iterations and report them to the relays.

$$\begin{aligned}\mu(\nu+1) &= \Big[\mu(\nu)-\xi_1(\nu)\Big(P_t-\sum_{m=0}^{M}\sum_{k\in\mathcal{Q}_m}\sum_{n=1}^{N}p_{kn}^{m}\Big)\Big]^{+},\\ \delta_n(\nu+1) &= \Big[\delta_n(\nu)-\xi_2(\nu)\Big(1-\sum_{m=0}^{M}\sum_{k\in\mathcal{Q}_m}\sum_{n=1}^{N}x_{kn}^{m}\Big)\Big]^{+},\forall n\end{aligned} \tag{4.12}$$

In (4.12), ν is the iteration index, and $\xi_1(\nu)$ and $\xi_2(\nu)$ are the step sizes at the ν-th iteration for updating μ and δ_n variables, respectively. The number of iterations can be optimized to reach fast convergence, by choosing suitable step sizes and initial values [10]. In this algorithm, the overhead of messages reported by relays is in the order of $O(NM)$ multiplied by the number of iterations, which is considerably lower than that of a centralized algorithm in the networks with high number of users.

Based on the above discussions, we note that DDRA reduces the computational burden of the BS and at the same time takes into account the QoS requirements of the users. Therefore, it provides a novel framework for distributed resource allocation in buffer-aided relay-assisted OFDMA networks.

4.4 Numerical Results

To evaluate the system performance, we have considered a system with cell radius equal to 1000 m, 3 relays and 20 subchannels. Relays are located at a distance of 2/3 cell radius from the BS, with uniform angular distance from each other. Antenna heights for the BS, relay and users are considered 15 m, 5 m and 1.5 m respectively and the path loss model is based on [61]. The noise power spectral density is assumed -174 dBm/Hz and BER requirements for users is considered 10^{-6}. The carrier frequency is 1900 MHz, and

the bandwidth for subchannels and the duration of time slots are considered equal to 15 kHz and 1 ms, respectively. Data packet arrivals at the BS buffers are based on Poisson process, with packet sizes equal to 1 kbits and the average packet interarrival time equal to 30 ms.

For the links from the BS or relays to users, Rayleigh channel model is used, while the links from the BS to relays are modeled with Rician channel with κ factor equal to 6 dB [62]. In the following, the results are presented in terms of system throughput as well as average queue sizes in the system. For baseline, we have used the Partial Proportional Fair (PPF) method proposed in [71] in which power is equally allocated over subchannels and relays are prompt, i.e., they transmit in a time subslot immediately after the reception subslot. We have adjusted PPF for our scenario by considering the availability of data in the users' queues in the BS; we call it Queue-Aware PPF (QAPPF), as it computes the achievable rates of users based on their queue size and channel conditions.

Fig. 4.3 displays the CDF of system average throughput in each time slot, in the case of 10 users in the system. It is observed that DDRA is able to provide higher throughput with higher probability, compared with QAPPF. This is due to the fact that the use of buffers in the relays brings more flexibility to resource allocation. Therefore, DDRA is able to utilize time diversity and allocate system power and subchannels efficiently. The jumps in the graph of DDRA is because of the fact that it is able to empty the buffers in some time slots. Then, when a new packet arrives at a buffer, it is transmitted completely; therefore, many transmissions occur in the units of a packet size, which are reflected as the jumps in the CDF graph. Also, we observe that this happens more in the case of higher system power, as it is possible to transmit with higher rates and have higher probabilities of empty buffers.

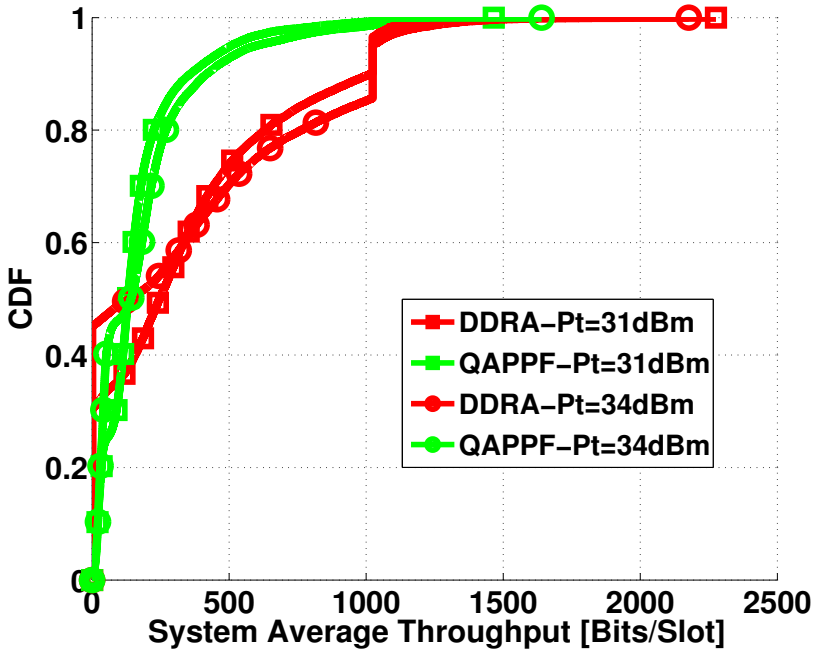


Figure 4.3: CDF of system average throughput in each time slot, K=10

To have a clearer picture, Fig. 4.4 demonstrates the average queue size over time, in this scenario. While the queue sizes grow unbounded with QAPPF, DDRA is able to keep the system queues stable. This is due to the fact that, according to (4.3), DDRA gives higher weights to the users with larger queue sizes, and using (4.9) and (4.11), it is able to allocate resources adaptively based on the queue sizes and channel conditions. As a result of queue stability, the average data departure rates of the queues are equal to their average arrival rates and therefore, as displayed in Fig. 4.3, DDRA is able to provide higher throughput.

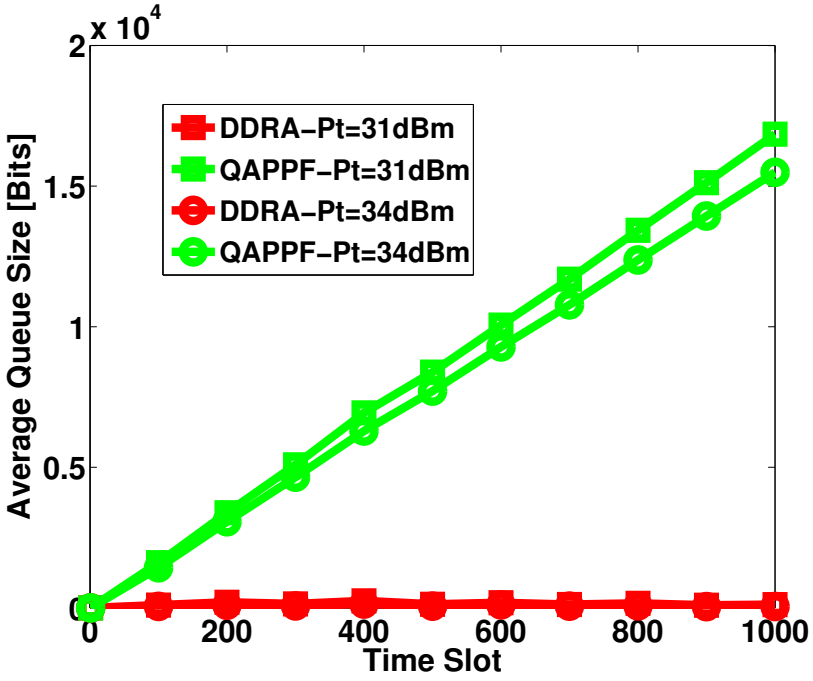


Figure 4.4: System average queue size over time, K=10

Next, we investigate the effect of increase in the number of users on the system throughput. Fig. 4.5 shows the average system throughput in each time slot. It is observed that as the number of users increases, DDRA leads to higher throughput. This is due to the fact that the total average data arrival rate in the system increases, and DDRA is able to stabilize the system and deliver the arrived data to their destinations. However, in the case that the system power is lower, its capacity is less; therefore, the capacity starts to get saturated when the number of users is more than 16. In this case, the system can not support the data arrival rates at the BS buffers and therefore, leads to lower throughput.

4.5 Summary

In this chapter, we have provided a novel framework for distributed resource allocation in a relay-assisted OFDMA network, with the assumption that relays are able to buffer data and

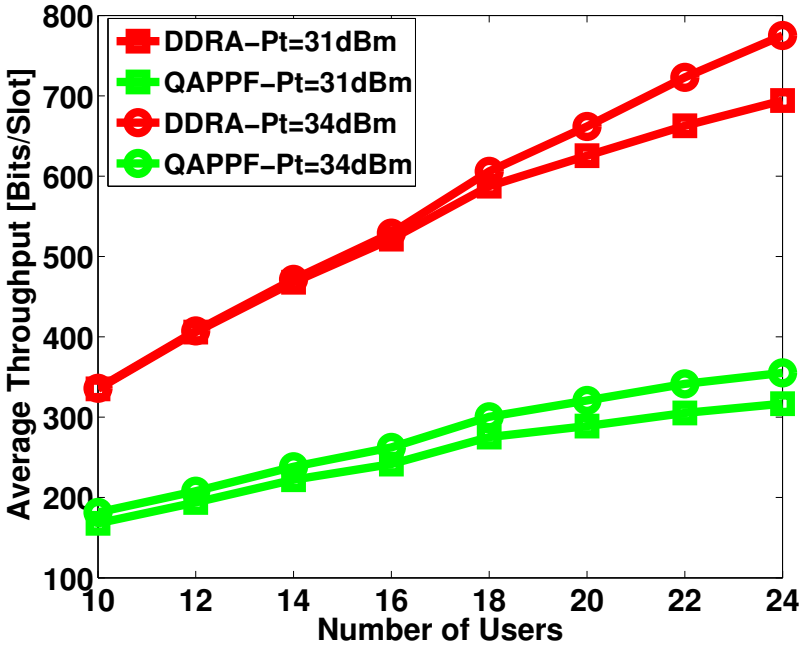


Figure 4.5: Effect of increase in the number of users on the system average throughput

transmit in a later time. By defining the links between the BS and relays as virtual users, we have presented a new perspective and showed the similarity of the system to a multicell network. We have formulated the resource allocation problem as a convex optimization problem and using dual decomposition, we have proposed the iterative DDRA algorithm, where each of the BS and relays solve their own problem based on some global variables and the information about the corresponding queue and channel states of their links. The closed form equations derived for power and subchannel allocation reveals the adaptive characteristic of our resource allocation algorithm to queue sizes and channel conditions of the users. Numerical results confirm that DDRA leads to significant improvement in the system performance in terms of average throughput and queue stability.

Chapter 5

Utility-Based Efficient Dynamic Distributed Resource Allocation

5.1 Introduction

In the previous chapter, we proposed a framework for distributed resource allocation in buffer-aided relay-assisted OFDMA networks, which provided insights for reducing the computational burden on the BS and the overhead of CSI signaling in the system. In this chapter, we study distributed resource allocation for BE users while taking into account more constraints that usually arise in practical systems.

In most of the existing studies on OFDMA relay networks, the resource allocation for BE users is formulated with the assumption that the buffers in the BS are infinitely backlogged and there are always data in the BS buffers. However, in practice, the data arrivals in the BS buffers are random and depending on the system objective and constraints, a data admission policy is needed to control the traffic flow into the network. In this regard, network operators take into account the average performance metrics such as average data admission and throughput as well as the average power constraints, in formulating the resource allocation problem. On the other hand, system hardware and technology standard specifications usually impose other constraints that need to be taken into account instantaneously in every time slot, to lead to a proper operation of the system.

To consider the above mentioned in the design of resource allocation algorithms, stochas-

tic network optimization is required. One of the useful and effective tools for addressing this class of problems is the well-known Lyapunov drift-plus-penalty policy which provides an efficient framework for transforming the stochastic problem into an instantaneous one to be solved in each time slot [51, 55][7]. However, using this policy without considering the constraints and challenges that arise in relay-assisted cellular networks can lead to unsatisfactory results. Specifically, in the drift-plus-penalty policy, the average of the variables is defined over infinite time horizon, which requires some considerations for achieving the desired objectives and satisfying the constraints in practical systems. Also, the lower power of the relays and the two-hop connection of some users necessitates careful attention in providing fair data admission for all the users. Moreover, it is of great interest to design low-complexity and efficient algorithms for deciding about the allocation of the system resources without incurring significant degradation in the system performance. To the best of our knowledge, none of the existing works on OFDMA relay networks has studied resource allocation with the above mentioned constraints altogether. This chapter aims at addressing these issues and filling the gaps.

In summary, we study low-complexity utility-based resource allocation for BE services in buffer-aided relay-assisted OFDMA networks, based on stochastic optimization framework presented in [55]. We consider the network utility as a function of average data admission of the users and aim at maximizing it subject to the long term and instantaneous constraints. In addition to considering several practical constraints altogether, the contributions of our work can be classified into two categories: one category is the identification of important parameters that should be taken into account in the instantaneous problem formulation. The other one is the design of low-complexity algorithms for solving the resource allocation subproblem. Specifically, the main contributions are as follows:

[7]In this chapter, when we refer to [55], we mean the Chapter 5 of that reference, unless otherwise specified.

- We identify the factors that need to be taken into account for adapting the Lyapunov drift-plus-penalty policy for relay-based cellular networks. In particular, we propose to consider an importance parameter for average power constraint, to satisfy that constraint in a reasonable time period for practical scenarios. Also, we propose to add extra weight for the BS-to-relays and relays-to-users links, in the cases that the fairness is also an objective in the utility-based data admission control.

- We aim at low-complexity algorithms for time slot, subchannel and power allocations and highlight the challenges even in such algorithms, due to the lack of a priori knowledge about the subchannel sets and total power usage of the BS and relays in each time slot. Then, we propose a low-complexity strategy for breaking the ties and making the interdependence tractable, which can be used in both centralized and distributed resource allocation implementations.

- We focus on distributed mechanism for resource allocation, and propose efficient and low-complexity algorithms for deciding about the type of time slot, subchannel sets of the nodes, and subchannel and power allocations to the links of the nodes. Based on that, we also present a low-complexity centralized mechanism which needs more signaling overhead and can be used as a benchmark.

- We take into account practical constraints such as average power, peak power as well as finite data availability and limited buffer space.

- Using extensive simulations, we demonstrate the effectiveness of the introduced parameters and verify the performance of the proposed algorithms. We observe that the distributed algorithm has very close performance to the centralized one and outperforms an existing centralized scheme proposed in [48].

The rest of this chapter is organized as follows. Section 5.2 describes the system model

and the stochastic problem formulations. In Section 5.3, we state the subproblems and challenges as well as the proposed parameters and algorithms. Numerical results are provided in Section 5.4, with conclusion finally presented in Section 5.5.

5.2 Preliminaries

In this section, we present the system model and the stochastic problem formulation. Then, we present the transformed version of the problem and introduce the virtual queues which make it possible to exploit the Lyapunov drift-plus-penalty policy in the next section. Hereafter, for easiness, we will use the term "drift-plus-penalty" instead of "Lyapunov drift-plus-penalty".

5.2.1 System Model

We consider the downlink of a single cell relay-assisted OFDMA network, as shown in Fig. 5.1. It is assumed that the users have BE services and therefore, they have not specific service requirements. In this regard, the system resources remained after serving the users with specific QoS requirements are considered for serving the BE users. We assume that each user is connected to either the BS or one of the relays, meaning that it receives service from only one of them. This is decided at the beginning of users' connection to the network and through handshaking procedures between the BS, relays and users, about the signal strengths that users can receive from the BS and relays. Users, relays and available subchannels are indexed respectively by $k \in \mathcal{K} = \{1, ..., K\}$, $m \in \mathcal{M} = \{1, ..., M\}$ and $n \in \mathcal{N} = \{1, ..., N\}$. We use the term "serving node" or simply "node" to refer to any of the BS or relays and we show the set of all nodes by $\mathcal{B} = \{0, 1, ..., M\}$, where $m = 0$ indicates the BS. Also, we use $\mathcal{K}_m$ to denote the set of users that have a direct link to node $m \in \mathcal{B}$. On the other hand, $m(k)$ is used to refer to the node directly connected to user k.

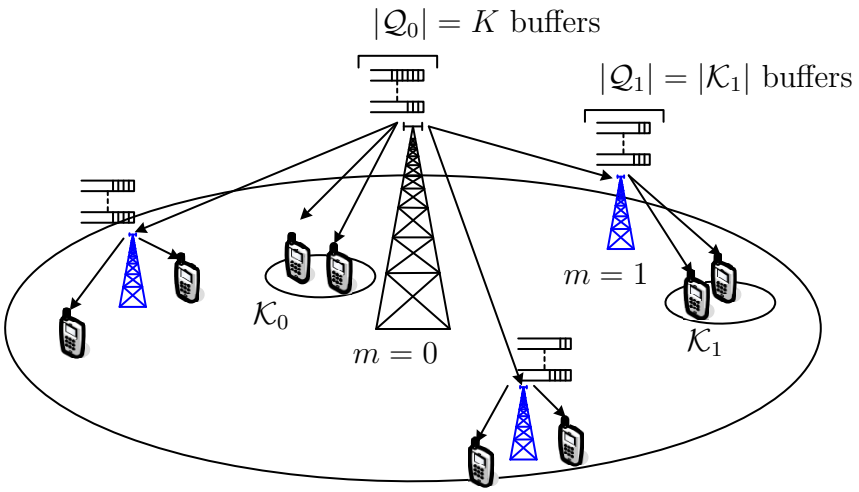


Figure 5.1: System model

We assume that time is divided into the units of slot, where a time slot can be type A or type B. In type A slots, the BS transmits to users directly connected to it, or to the relays; in type B slots, the BS and relays can transmit to only the users connected to them and therefore, there is no transmission from the BS to relays. This transmission format is based on LTE-A with type 1 relays where BS-to-relays transmissions and relays-to-users transmissions use the same bandwidth but over different time slots, in order to prevent the interference between transmit and receive antennas.

We assume that the MAC layers of the BS and relays are equipped with buffers, where the BS has one for each user but every relay has one for each of the users connected to it. We denote the set of the users that have a buffer in node $m \in \mathcal{B}$ by $\mathcal{Q}_m$; therefore, we have $\mathcal{Q}_0 = \mathcal{K}$ and $\mathcal{Q}_m = \mathcal{K}_m, \forall m \in \mathcal{M}$. These notations are defined to make the formulations and algorithms shorter. The data admitted into a BS buffer, are queued until transmission to the corresponding direct user, or the corresponding buffer in the relay connected to the user. Similarly, data arrived in relays' buffers, are queued until transmission to their users.

Note that in the following, when we use the term "the link of user k from node m", we mean "the link that serves the queue of user k in node m", which might be a *direct*

link between the BS and a user, a *feeder* link between the BS and a relay or an *access* link between a relay and a user connected to it. We use $e_{kn}^{m}(t)$ for the link of user k from node m, to denote the channel gain-to-noise ratio at the receiver side on subchannel n. It is assumed that the channel conditions of the links vary over time and frequency, but remain constant during one time slot and over one subchannel. For simplicity, we will remove the (t) argument from the variables and imply them in a general transmission incident. We assume that the BS and relays use M-ary QAM modulation for their transmissions, and therefore, the achievable transmission rate on the link of user k from node m on subchannel n can be computed as follows [63]:

$$r_{kn}^{m} = B \log_2 \left(1 + \frac{p_n^m e_{kn}^m}{\Gamma_k}\right), \tag{5.1}$$

where B is the bandwidth of a subchannel. Γ_k is the SNR gap due to the limited number of coding and modulation schemes and is related to the bit error rate of user k (BER_k), through equation $\Gamma_k = -\frac{\ln(5BER_k)}{1.5}$ [63]. p_n^m denotes the power allocated by node m on subchannel n. We indicate the total power used by node m as $P_m = \sum_{n=1}^{N} p_n^m$. Using (5.1), the total transmission rate on the link of user k from node m can be written as $r_k^m = \sum_{n=1}^{N} x_{kn}^m r_{kn}^m$, where $x_{kn}^m \in \{0, 1\}$ denotes the subchannel allocation indicator which will be one if subchannel n is used for transmission on the link of user k from node m, and zero otherwise. Note that for any n, in type A time slot, x_{kn}^m should be set to zero for $m \in \mathcal{M}, k \in \mathcal{K}_m$ and in type B time slot, x_{kn}^0 should be set to zero for $k \in \mathcal{K} - \mathcal{K}_0$.

In each time slot, a resource allocation policy determines the type of time slot, and subchannel and power allocations for the different links of the system. Based on that, the BS and relays transmit data from their queues and at the end, the queue sizes are updated

as follows:

$$Q_k^m(t+1) = \min[L_k^m, \max[Q_k^m(t) - Tr_k^m(t), 0] + a_k^m(t)], \forall m \in \mathcal{B}, k \in \mathcal{Q}_m \tag{5.2}$$

where T is time slot duration and L_k^m and $Q_k^m(t)$ respectively denote the buffer capacity of user k in node m and the size of data queued in it in time slot t. Data arrival processes into relay buffers are in fact the departure processes from the BS and therefore, we have $a_k^m(t) = \min[Q_k^0(t), Tr_k^0(t)], \forall m \in \mathcal{M}, k \in \mathcal{K}_m$. On the other hand, the arrival processes at the BS buffers are managed through a data admission control policy in the MAC layer, which decides to admit or reject data from the queues of the top layer buffers in the BS. The size of the queues in top layer are updated as follows:

$$Y_k(t+1) = \min[J_k, \max[Y_k(t) - a_k^0(t), 0] + A_k(t)], \ \forall k \in \mathcal{K} \tag{5.3}$$

where J_k and $Y_k(t)$ respectively denote the buffer capacity of user k in the top layer of the BS and the size of data queued in it in time slot t. $A_k(t)$ is the amount of data arrived in time slot t for user k, according to an exogenous stochastic process. We assume that due to processing limitations, an upper bound of $\hat{a}$ is imposed on the amount of data admitted into the MAC layer buffers and therefore, we have $a_k^0(t) \leq \hat{a}, \forall k, t$.

5.2.2 Stochastic Problem Formulation

We note that in a realistic scenario, the BS buffers are not infinitely backlogged and are fed by stochastic data arrivals. This makes it necessary to take into account the queue dynamics, in the design of resource allocation algorithms, in addition to the randomness caused by the wireless channels. Therefore, the average performance metrics become important for network operators, and specially, the average throughput and average power

constraints are the issues that need to be managed. In the following, we will explain these in more detail.

According to [55, Chapter 4], we define the time average expectation of a stochastic variable $v(t)$ as:

$$\overline{v} = \lim_{\tau \to \infty} \frac{1}{\tau} \sum_{t=0}^{\tau-1} E[v(t)] \tag{5.4}$$

Considering the above mentioned, we aim at controlling the data traffic and resource allocation, by addressing the following stochastic optimization problem:

$$\max_{\boldsymbol{a}^0, \boldsymbol{x}, \boldsymbol{p}} \sum_{k=0}^{K} \mathcal{U}(\overline{a}_k^0), \tag{5.5a}$$

$$\text{s.t.} \quad C1 : \overline{P}_m \leq P_m^{av}, \quad \forall m \in \mathcal{B}, \tag{5.5b}$$

$$C2 : r_k^m(t)T \leq Q_k^m(t), \quad \forall m \in \mathcal{B}, k \in \mathcal{Q}_m, \tag{5.5c}$$

$$C3 : r_k^0(t)T \leq (L_k^m - Q_k^m(t)), \quad \forall m \in \mathcal{M}, k \in \mathcal{K}_m, \tag{5.5d}$$

$$C4 : r_{kn}^m(t) \leq B\hat{s}, \quad \forall m \in \mathcal{B}, k \in \mathcal{Q}_m, \forall n, \tag{5.5e}$$

$$C5 : a_k^0(t) \leq Y_k(t), \quad \forall k \in \mathcal{K}, \tag{5.5f}$$

$$C6 : a_k^0(t) \leq \min[\hat{a}, L_k^0 - Q_k^0(t)], \quad \forall k \in \mathcal{K}, \tag{5.5g}$$

$$C7 : P_m(t) \leq \hat{P}_m, \quad \forall m \in \mathcal{B}, \tag{5.5h}$$

$$C8 : \sum_{m \in \mathcal{B}} \sum_{k \in \mathcal{Q}_m} x_{kn}^m(t) \leq 1, \quad \forall n \in \mathcal{N}, \tag{5.5i}$$

$$C9 : \{x_{kn}^m(t)\} \text{ comply to the transmission rules of either type A or B slot} \tag{5.5j}$$

where $\mathcal{U}$ is the utility function and $C1$ is to limit the average power consumption of each node. $C2$ shows that a finite amount of data can be transmitted from each queue and $C3$ is to prevent the incidents of more transmissions to relay buffers than they can accommodate. $C4$ indicates the limit on the availability of modulation schemes, where $\hat{s}$ is the spectral

efficiency of the highest order modulation in the system; considering it helps in controlling the power allocation and preventing overflows from relays' buffers (This will be explained clearly in Section 5.3). $C5, C6$ respectively show the limit on the availability of data in the top layer buffers of the BS and the limit on the data admissions. $C7$ indicates the maximum instantaneous power, $\hat{P}_m$, that node m can use for transmissions, $C8$ shows that each subchannel can be allocated to only one link, and $C9$ is to use the feasible values for subchannel allocation variables $\{x_{kn}^m\}$.

The utility function in (5.5a) makes it possible to control the data admission of the BE users based on the objective of the network operator. For example for maximizing the total throughput, $\mathcal{U}(z) = z$ can be used or for providing proportional fairness, $\mathcal{U}(z) = \log(z)$ can be considered. We assume that $\mathcal{U}(z)$ is a concave and continuous function of z.

We note that the problem (5.5) has two types of constraints. While $C1$ needs to be satisfied over long time, $C2-C9$ state the constraints that must be met in each time slot. In particular, P_m^{av} is different from $\hat{P}_m$, as the former can be set to limit the power consumption costs or the circuit heating but the latter is imposed by the system hardware (such as power amplifiers' linear operation characteristics or maximum available instantaneous power) and is larger than P_m^{av}.

5.2.3 Transformed Problem and Virtual Queues

We note that the objective in problem (5.5) is a function of time average of users' data admission rate. Similar to [55], let define auxiliary variables $0 \leq \gamma_k(t) \leq \hat{a}, k = 1, ..., K$, corresponding to each $a_k^0(t), k = 1, ..., K$. Then, the problem (5.5) can be transformed into

the following problem, in which the objective is time average of a function:

$$\max_{\boldsymbol{a}^0,\boldsymbol{x},\boldsymbol{p},\boldsymbol{\gamma}} \sum_{k=0}^{K} \overline{\mathcal{U}(\gamma_k)}, \tag{5.6a}$$

$$\text{s.t.} \quad C1 - C9, \tag{5.6b}$$

$$C10 : \overline{\gamma}_k \leq \overline{a}_k^0, \forall k \in \mathcal{K}, \tag{5.6c}$$

$$C11 : 0 \leq \gamma_k \leq \hat{a} \tag{5.6d}$$

We also define the virtual power queues $Z_m(t)$ and virtual auxiliary queues $G_k(t)$, respectively corresponding to the constraints $C1$ and $C10$, with the following update equations:

$$Z_m(t+1) = \max[Z_m(t) + P_m(t) - P_m^{av}, 0], \qquad \forall m \in \mathcal{B} \tag{5.7a}$$

$$G_k(t+1) = \max[G_k(t) + \gamma_k(t) - a_k^0(t), 0], \qquad \forall k \in \mathcal{K} \tag{5.7b}$$

$Z_m(t)$ is in fact like a queue that has a stochastic arrival process $P_m(t)$ and a deterministic constant service process P_m^{av}. Similarly, $G_k(t)$, acts like a queue that has the stochastic arrival process $\gamma_k(t)$ and stochastic service process $a_k^0(t)$. Note that over time, these virtual queues track the difference of the instantaneous values of the variables on both sides of the constraints $C1$ and $C10$. The idea behind using these queues is the fact that satisfying the constraints $C1$ and $C10$ is equivalent to stabilizing these virtual queues[8] which can be performed by using drift-plus-penalty policy (of course, assuming that the original problem is feasible) [55]. Based on the above mentioned, we are able now to study the instantaneous problem and the algorithms for solving it, which will be presented in the next Section.

[8]Recall that when a queue is stable, it means that its average service rate is larger than its average arrival rate

5.3 Cross Layer Traffic Control and Resource Allocation

In this section, we first describe the instantaneous problem to be addressed in each time slot and propose some parameters that need to be included in it to make it suitable for relay-assisted cellular networks. Then, we present the data admission subproblem and discuss the factors influencing it. After that, we highlight the issues in solving the resource allocation subproblem and propose a low-complexity strategy to address them. Then, we provide a low-complexity distributed algorithm which uses the proposed strategy through several steps for deciding about the allocation of time slots, power and subchannels. Finally we present a low-complexity centralized algorithm, based on the distributed one and describe the modifications needed.

5.3.1 Instantaneous Problem

To address the problem (5.6), we define the "instantaneous" problem in time slot t, based on the drift-plus-penalty policy [55], as follows:

$$\max_{\boldsymbol{a}^0,\boldsymbol{x},\boldsymbol{p},\boldsymbol{\gamma}} \quad V\sum_{k=0}^{K}\mathcal{U}(\gamma_k(t)) + \sum_{k=0}^{K} G_k(t)[a_k^0(t)-\gamma_k(t)] + I\sum_{m=0}^{M} Z_m(t)[P_m^{av}-P_m(t)] + \sum_{m=0}^{M}\sum_{k\in\mathcal{Q}_m}(Q_k^m(t)+\rho_k^m W_e)[r_k^m(t)T-a_k^m(t)], \tag{5.8a}$$

$$\text{s.t.} \quad C2-C9, C11 \tag{5.8b}$$

where V is the value that can be given to the objective (5.6a), and by that we can trade off higher objective to larger queue sizes[55]. I, ρ_k^m and W_e are the parameters that we propose, to adapt the drift-plus-penalty policy to relay-assisted cellular networks. I is the importance factor that we give to average power constraint, through which we can prevent the continuous growth of the virtual power queues and consequently, we can meet the average power constraints in shorter time. W_e is an extra positive weight, that can be given for the feeder links from the BS to relays and the access links from relays to users, in the cases that the fair admission of users' data is of our concern. $\rho_k^m \in \{0, 1\}$ is the indicator to specify the conditions and the queues that can exploit W_e; In particular, it will be zero unless when the fair data admission is desired and the corresponding queue (either in the BS or in a relay) belongs to an indirect user, i.e., $k \in \mathcal{K} - \mathcal{K}_0, m \in \mathcal{B}$. Proposition of I, ρ_k^m and W_e is one of the main contributions of this chapter and will be discussed later.

It is observed from (5.8a) that the instantaneous objective in each time slot, includes four terms: The first term corresponds to the long term objective (5.6a) and the rest correspond to serving the actual queues and stabilizing the virtual queues (to meet the constraints $C1$ and $C10$). We note that due to the limited buffer capacities, the actual queues of the system are always stable. However, using drift-plus-penalty policy provides a useful framework for channel-and-queue-aware resource allocation which takes into account both channel states and the data availability in the system's actual queues. It also makes it possible to stabilize virtual power queues.

Note that considering V, I, ρ_k^m and W_e in (5.8a) is to facilitate reaching our purposes for system utility and constraints, in different scenarios; otherwise neglecting them is in fact like setting them based on a fixed scenario (i.e., V=I=1 and W_e=0) which would lower the usefulness of the drift-plus-penalty policy. Note that V and I can be tuned easily by considering the range of values for weighted rates in (5.8a), affected by packet sizes and

transmission rates. On the other hand, ρ_k^m can be easily set as stated above, when fairness is an objective; then, W_e can be tuned by increasing its value from zero towards the values in the range of queue sizes in relays, depending on how much we trade off between data admission for the direct and indirect users. These will get clearer in the following, when we discuss their effects.

Similar to [55], by rearranging the terms in the instantaneous problem, it can be divided into three instantaneous Subproblems (SPs) which will be presented in the next subsections.

5.3.2 Traffic Control and Data Admission

The first subproblem of (5.8) is related to the auxiliary variables as follows:

$SP1$ (Auxiliary Variables Subproblem):

$$\max_{\boldsymbol{\gamma}} \sum_{k=1}^{K} \Big(V\mathcal{U}(\gamma_k(t)) - G_k(t)\gamma_k(t) \Big), \tag{5.9a}$$

$$\text{s.t.} \;\; 0 \leq \gamma_k(t) \leq \hat{a}, \forall k \in \mathcal{K} \tag{5.9b}$$

SP1, (5.2), (5.3), (5.7b) and the following flow control subproblem, affect the admission of the data into the users' buffers in the BS:

$SP2$ (Flow Control Subproblem):

$$\max_{\boldsymbol{a}^0} \sum_{k=1}^{K} \Big(G_k(t) - Q_k^0(t) \Big) a_k^0(t), \tag{5.10a}$$

$$\text{s.t.} \;\; 0 \leq a_k^0(t) \leq \hat{a}, \forall k \in \mathcal{K}, \tag{5.10b}$$

$$a_k^0(t) \leq \min[Y_k(t), L_k^0 - Q_k^0(t)], \forall k \in \mathcal{K} \tag{5.10c}$$

It is observed that the SP1 and SP2 are simple convex problems; all of their variables are accessible to the BS and therefore, the BS can solve them easily. As an example, for the proportional fairness, i.e., when $\mathcal{U}(\gamma_k) = \log(\gamma_k)$, after taking the derivatives, the BS can determine the optimal auxiliary variables in time slot t through $\frac{1}{\gamma_k(t)} = \frac{G_k(t)}{V} \Longrightarrow \gamma_k(t) = \min(\hat{a}, \frac{V}{G_k(t)}), \forall k$.

On the other hand, by solving the flow control subproblem, the BS can determine the optimal data admissions in time slot t as:

$$a_k^0(t) = \begin{cases} \min[Y_k(t), \min[\hat{a}, L_k^0 - Q_k^0(t)]] & , Q_k^0(t) \leq G_k(t) \\ 0 & , \text{otherwise} \end{cases} \tag{5.11}$$

Note that, based on the above mentioned, whenever the size of an actual queue in the BS, $Q_k^0(t)$, is larger than the virtual queue $G_k(t)$, no data is admitted into the corresponding BS buffer. This can happen for several time slots, in buffer-aided relay-assisted cellular networks, due to the admission of a large packet and low service rate of the queue of an indirect user in the BS (which is caused by the differential backlog terms described in the next subsection). Consequently, even using a utility function with fairness property and large value for parameter V does not necessarily lead to fair data admission, and therefore, more considerations are needed in resource allocation for serving the queues. This is the motivation for proposing ρ_k^m and the extra weight W_e (explained clearly later, in Remark 1), which help to improve the fair data admission for indirect users.

5.3.3 Resource Allocation Challenges

By substituting (5.1) in (5.8a) and removing the constant terms, the last and most important subproblem, which is to decide about the time slot, subchannel and power allocations

can be stated as

$SP3$ (Resource Allocation Subproblem):

$$\max_{\boldsymbol{p},\boldsymbol{x}} \sum_{n=1}^{N}\sum_{m=0}^{M}\Big(\sum_{k\in\mathcal{Q}_m} BTw_k^m(t)x_{kn}^m(t)\log_2(1+\frac{p_n^m(t)e_{kn}^m(t)}{\Gamma_k}) - IZ_m(t)p_n^m(t)\Big), \tag{5.12a}$$

$$\text{s.t.} C2-C4, C7-C9 \tag{5.12b}$$

where $w_k^0(t)=Q_k^0(t), k\in\mathcal{K}_0$ (weights for the direct links from the BS to its users; recall that ρ_k^0 is always equal to 0 for these links), $w_k^m(t)=Q_k^m(t)+\rho_k^m W_e, m\in\mathcal{M}, k\in\mathcal{K}_m$ (weights for the access links from relays to their users), and $w_k^0(t)=Q_k^0(t)-Q_k^{m(k)}(t)+\rho_k^0 W_e, \forall k\in\mathcal{Q}_0-\mathcal{K}_0$ (weights for the feeder links of indirect users from the BS to relays). The differential backlog term $Q_k^0(t)-Q_k^{m(k)}(t)$ in the weight of a feeder link is resulted by switching the sums and considering the fact that for the buffers of the relays, the arrivals are upper bounded by the transmission rates from the BS to relays, i.e., $a_k^m(t)\leq r_k^0(t), \forall m\in\mathcal{M}, k\in\mathcal{Q}_m$[9]. As explained in the previous subsection, and later in the Remark 1, these differential backlog terms lead to unfair data admission for indirect users, and their effect can be reduced by using $\rho_k^0 W_e$ terms.

We note that SP3 is a mixed integer and nonlinear programming and needs an exhaustive search to find its optimum solution. One common approach for these types of problems is to relax the subchannel allocation variables x_{kn}^m whenever this relaxation converts the problem into a convex one. Then, using the dual decomposition, optimal solution can be found, if the duality gap is zero. This approach was used in the previous chapter, based on which we proposed a dynamic distributed resource allocation. However, in this chapter, due to the finite data and limited buffer capacity constraints, i.e., $C2$ and $C3$, the resulted problem after relaxation of x_{kn}^m variables will be non-convex. In addition, we note that in

[9]Note that the objective function in the MW policy, utilized in the previous chapters, is in fact the special case of that in SP3 where the virtual power queues are not considered and $W_e=0$.

the previous chapter, there was only one power constraint for the whole system which made it possible to have high convergence speed for the proposed algorithm. In a more realistic system like the one we consider in the current chapter, even if we remove the constraints $C2, C3$ and relax x_{kn}^m variables to make it convex, a dual based iterative algorithm will need many iterations and a long time to converge, due to the separate power constraints of nodes. This is not suitable for use in practical scenarios where each time slot is in the order of a millisecond and the resource allocation decision needs to be made in a small fraction of time.

Due to the above mentioned, we aim at low-complexity suboptimal algorithms which can be easily implemented in practical systems. For this purpose we consider equal power allocation on subchannels and allocate them in a greedy way, based on the queue sizes and achievable transmission rates of the links.

However, even considering equal power distribution on subchannels and computing the achievable transmission rates of the links is not trivial here, due to the following two issues:

a) **Unknown number of subchannels for each node.** For deciding about the allocation of subchannels, we need to know the achievable transmission rates of the links on the subchannels, and for that we need to know the power allocations on the subchannels. However, before subchannel allocation, it is not clear how many subchannels will be allocated to the BS and relays, and consequently it is not known on how many subchannels their total powers will be distributed equally.

b) **Unknown total powers to be used by each node.** The total powers used by each of the BS and relays, need to satisfy the average and peak power constraints. This is controlled in SP3 through the objective function, which is sum of the increasing and decreasing functions of power, and constraint $C7$. Based on that, the total power used by each node can vary in each time slot between zero and its peak power,

Algorithm 5.1 Subchannel and Power Allocation Strategy (SPAS)

- Assume the number of subchannels to be assigned to node m, $\tilde{N}_m$, proportional to the number of its queues.
- Assume the power each node will use on the subchannels, will be based on peak power, and equal to $\frac{\hat{P}_m}{\tilde{N}_m}$.
- Compute the achievable transmission rates of the links based on the channel conditions and assumed powers.
- Determine the type of time slot and allocate the subchannels, based on the actual queue sizes and the achievable transmission rates.
- Considering the the size of actual and virtual queues, adjust the total power each node can use and distribute it equally on the subchannels assigned to it in the previous steps.

depending on the subchannel allocations and the sizes of the corresponding virtual queues. Therefore, even if we make an assumption on the number of subchannels to be used by each node, it is not clear that how much total power will be distributed equally on them.

To address the stated issues, we propose a low-complexity suboptimal strategy for breaking the interdependence of power allocations and subchannel assignments, as shown in Subchannel and Power Allocation Strategy (SPAS).

Note that with SPAS, the total transmission power of each node is assumed equal to the peak value, at the beginning. Based on that, subchannels are allocated and at the end, the total power is adjusted. The reason for this is clarified later in Remark 2.

SPAS can be utilized in a centralized or distributed way, with some modifications. In the following we will present the distributed implementation, as it is of more interest to the research and industrial bodies; later based on that, we will describe the centralized resource allocation, which can be used as a benchmark for the proposed distributed one.

5.3.4 Efficient Dynamic Distributed Resource Allocation

In this subsection, we propose the EDDRA method which performs resource allocation in each time slot, through four steps. In the first step, every node reports estimates of its subchannel demands to the BS and based on them, the BS decides about the type of time slot. In the second step, the BS determines and reports the subchannel sets that each of the BS and relays can use. Then, in the third and forth steps, in a distributed way, each node first assigns the subchannels to its users and then, it adjusts the total power it can distribute over its subchannels.

Step 1) Slot Type Determination(STD). At the end of each time slot, first the BS needs to specify the type of the next slot. For this, relays report an estimate of their average demand for each subchannel, to the BS. These demands are computed based on the assumptions on subchannel numbers they can get and the total power they can use. Then, the BS uses the reported demand estimates from relays as well as its own demands, to estimate the system's total demands for type A and type B slots, and based on that, decides about the type of the next time slot. This is outlined in STD algorithm and the details are described in the following.

Based on SP3, we define the estimated average demand of node $m \in \mathcal{M}$ on subchannel n as

$$D_n^m = \frac{1}{|\mathcal{K}_m|} \sum_{k \in \mathcal{K}_m} w_k^m \tilde{r}_{kn}^m, \forall m \in \mathcal{M}, \forall n \in \mathcal{N} \tag{5.13}$$

where $\tilde{r}_{kn}^m = \log_2(1 + \frac{\hat{P}_m e_{kn}^m}{\tilde{N}_m \Gamma_k})$ is the estimated transmission rate of the link of user k from node m on subchannel n. It is computed in node m, $m \in \mathcal{M}$, assuming that the number of subchannels it will get, $\tilde{N}_m$, is proportional to the ratio of its number of queues ($|\mathcal{K}_m|$) and the total number of queues that can be considered for service in type B slot ($\sum_{m \in \mathcal{B}} |K_m| =$

Algorithm 5.2 Slot Type Determination (STD)

1: Each relay $m \in \mathcal{M}$ reports to BS, its estimated average demands on all subchannels, i.e., $D_n^m, \forall n \in \mathcal{N}$.
2: BS estimates the total demand of each relay m as $D^m = |\mathcal{K}_m| \sum_{n=0}^{N} D_n^m, m \in \mathcal{M}$
3: BS estimates its own demand for type A slot as $D^{0a} = K \sum_{n=0}^{N} D_n^{0a}$
4: BS estimates its own demand for type B slot as $D^{0b} = |\mathcal{K}_0| \sum_{n=0}^{N} D_n^{0b}$
5: BS estimates the total demand for type A slot as $D^A = D^{0a}$,
and for type B slot as $D^B = D^{0b} + \sum_{m=1}^{M} D^m$
6: **if** $D^A > D^B$
7: BS sets the type of slot to A
8: **else**
9: BS sets the type of slot to B
10: **end if**

K); i.e.,

$$\tilde{N}_m = N \frac{|\mathcal{K}_m|}{K}, \forall m \in \mathcal{B}, \tag{5.14}$$

Since the BS knows the number of queues in each relay, it can easily estimate their total demand as in line 2 of STD algorithm. For itself, the BS needs to compute separate demands for type A and type B slots. Noting that in type B slots, it can only transmit to its direct users while sharing subchannels with relays, its average demands are computed similar to relays and based on the weights and rates of the links of direct users (assuming the transmission power on each subchannel equal to $\frac{\hat{P}_0}{\tilde{N}_0}, \tilde{N}_0 = N\frac{|\mathcal{K}_0|}{K}$), i.e., $D_n^{0b} = \frac{1}{|\mathcal{K}_0|} \sum_{k \in \mathcal{K}_0} w_k^0 \log_2(1 + \frac{\hat{P}_0 e_{kn}^0}{\tilde{N}_0 \Gamma_k})$.

On the other hand, since in type A slot, only the BS can transmit and all the queues in the BS (including those of indirect users) can be served using all the subchannels, its total demand is computed based on the weights and rates of all of its links and assuming $\frac{\hat{P}_0}{N}$ power on each subchannel, i.e., $D_n^{0a} = \frac{1}{K} \sum_{k \in \mathcal{K}} w_k^0 \log_2(1 + \frac{\hat{P}_0 e_{kn}^0}{N \Gamma_k})$.

Note that for computing the demands in the BS, it needs to know the queue sizes of the relays as well (to be used in the weights of the feeder links from the BS to relays). For

this purpose, relays also report the information about their modified queue sizes, to the BS. Considering the fact that in each time slot, at most N different queues can be served, the maximum number of modified queue sizes in relays is $\min(K-|\mathcal{K}_0|, N))$. Therefore, in EDDRA, the total overhead of signaling about the demands and the modified queue sizes is of $O(\min(K-|\mathcal{K}_0|, N)+MN)$.

Remark 1: Here we explain the reason for using $\rho_k^m W_e$ in the weights of the links of indirect users from the BS and relays. Without that, due to the low powers of relays and low transmission rates, their demands would not be comparable to the demands of BS for direct users, unless the queues in relays grew large. On the other hand, for the links serving the queues of indirect users in the BS, we would have $w_k^0 = Q_k^0 - Q_k^{m(k)}, \forall k \in \mathcal{K}-\mathcal{K}_0$. As a result, these would not have enough impact on computing the average demands for indirect users and providing service for them (in the cases that the queue sizes of an indirect user in the BS and relay have the same size, $Q_k^0 = Q_k^{m(k)}$, the impact would be zero). Consequently, the queues of indirect users in the BS would usually have larger sizes than the queues of direct users and therefore, data admission would be less for them. This would degrade the usefulness of drift-plus-penalty for cellular networks, because fairness is usually one of the main concerns in these networks. To prevent that, W_e should be applied to compensate for the effect of low power of relays on the demands of access links and the effect of differential-backlog-based weights on the demands of feeder links. Similar effect holds also in the subchannel sets determination and subchannel allocation steps which will be described later.

Step 2) Subchannel Sets Determination (SSD). We note that due to sharing subchannels in type B time slot among all the links, the resource allocation for the links of different nodes are tied together which is reflected in (5.12). In this step, the goal is to break this tie and specify the subchannel sets to be used for transmissions from the BS and

Algorithm 5.3 Subchannel Sets Determination (SSD)

1: **if** the slot type is A
2: BS determines the subchannels sets as $\mathcal{N}_0 = \mathcal{N}$ and $\mathcal{N}_m = \emptyset, m \in \mathcal{M}$
3: **else**
4: BS specifies subchannel sets, based on the relays' demands as well as its own, as follows:
5: Set $D_n^0 = D_n^{0b}$
6: Set $\hat{N}_m = \lceil N \frac{|\mathcal{K}_m| \sum_{n=0}^{N} D_n^m}{\sum_{m=0}^{M} \sum_{n=0}^{N} |\mathcal{K}_m| D_n^m} \rceil, \forall m \in \mathcal{B}$
7: Initialize $\mathcal{N}' = \mathcal{N}, \mathcal{B}' = \mathcal{B}, \mathcal{N}_m = \emptyset, \forall m \in \mathcal{B}'$
8: **while** $\mathcal{N}' \neq \emptyset$ and $\mathcal{B}' \neq \emptyset$
9: find $(m^*, n^*) = \arg \max_{m \in \mathcal{B}', n \in \mathcal{N}'} D_n^m$
10: $\mathcal{N}_{m^*} = \mathcal{N}_{m^*} \cup \{n^*\}$
11: $\mathcal{N}' = \mathcal{N}' - \{n^*\}$
12: **if** $|\mathcal{N}_{m^*}| = \hat{N}_{m^*}$
13: $\mathcal{B}' = \mathcal{B}' - m^*$
14: **end if**
15: **end while**
16: **end if**
17: BS notifies relays about their subchannel sets

relays. This allows to have the resource allocation in a distributed manner at each node.

For the above purpose, when the slot is decided to be type A, the BS notifies the relays about it and they know they have no transmissions. In the case of type B slot, the BS determines the subchannel sets of the relays and notifies them to transmit on them. SSD algorithm shows the whole procedure in detail. Since in the type B slots, the BS can only transmit to the direct users, line 5 of the algorithm defines its demands based on the estimations for type B slot, explained before. Line 6 sets $\hat{N}_m$, as the upper bound for the number of the subchannels that each node can get, and the next lines assign the subchannels to the nodes that have not reached their limit for subchannel numbers and have higher average demands on the subchannels.

Note that, setting the limit $\hat{N}_m$ for the size of the subchannel set for node m is to prevent the subchannel allocation far more than it needs. For example, a relay might

have only one user with high *average* demands on the subchannels while another relay with several users might have a little lower *average* demands on the subchannels. In such a case, without considering the total number of users and the limit for subchannel set sizes, the relay with single user would overshadow the other relay, in all the iterations of subchannel assignments through line 9.

The computational complexity of the SSD algorithm is of $O((M+1)N^2)$, which is obtained by ignoring the insignificant computations and considering the number of iterations needed for performing line 9.

Step 3) Subchannel Allocation (SA). After the determination of the subchannel sets, the resource allocation subproblem (5.12) can be further decomposed into separate **subsubproblems**, as follows:

SSP (Resource Allocation Subsubproblems):

$$\max_{\boldsymbol{p},\boldsymbol{x}} \sum_{n\in\mathcal{N}_m}\sum_{k\in\mathcal{Q}_m}\Big(BTx_{kn}^m(t)w_k^m(t)\log_2(1+p_n^m(t)e_{kn}^m(t))\Big) - \sum_{n\in\mathcal{N}_m} IZ_m(t)p_n^m(t), \forall m\in\mathcal{B}, \tag{5.15a}$$

$$\text{s.t.} C2-C4, C7-C8 \tag{5.15b}$$

where each node knows its set of subchannels and can decide individually about allocating them to its own links, considering the related subset of the constraints $C2-C4, C7-C8$. For this purpose, following the SPAS strategy, we propose to have subchannel allocations by each node based on using $\frac{\hat{P}_m}{|\mathcal{N}_m|}$ (i.e., assuming $Z_m(t)=0$) for computing the achievable transmission rates. Then, in the power adjustment step, considering the real value of $Z_m(t)$, each node can decide about the total power it should use and distribute it on its subchannels.

Noting that the BS has more constraints than other nodes, (the constraint $C3$ is only

enforced on the feeder links from the BS, which is to prevent transmitting data to the relays more than the empty buffer spaces), we provide the subchannel allocation by the BS and then, we explain its use for the relays. SA algorithm shows the details in allocating the subchannels by the BS. The procedure is done in $|\mathcal{N}_0|$ steps. In each step, the weights of the links and the resulted demands are computed, and the pair of subchannel and queue with the corresponding highest demand is determined. This is done in an iterative way and in each iteration, one subchannel is allocated and the affected queue sizes are updated virtually. Since the actual queue sizes, Q_k^m, are only updated at the end of transmission intervals, we have used q_k^m variables to prevent ambiguity about the updating during the algorithm iterations. Note that these updates are done to meet constraint $C2$ and $C3$. Line 6 is for applying the extra weight W_e, described before. However, before adding it, by comparing the queue size with $BT\hat{s}$, we make sure that there are enough data such that it can utilize the channel completely, in case subchannel is assigned. Line 7 is to meet the constraint $C3$ and prevent overflow, by giving a negative weight in case the remaining empty space in a relay buffer is less than the possible maximum transmission size on a subchannel. If a link gets negative weight, then, it will not be considered for subchannel allocation and this will prevent transmitting data to the corresponding relay buffer.

Remark 2: Note that the rate computations in SA are based on the assumption of equal distribution of peak powers on the subchannel sets. This way we will be sure that when in step 4, the total power is adjusted (which certainly will be equal or less than the peak power), the transmission rates for each link will be less than the amounts considered in SA algorithm, and therefore, the constraints $C2$ and $C3$ will not be violated.

In type B time slots, in parallel to the BS, any relay also uses the SA algorithm with the difference that all the superscripts/subscript 0 are replaced by the corresponding m. Note that in this case, we have $\mathcal{Q}' = \mathcal{K}_m$ and $\mathcal{Q}' - \mathcal{K}_m = \emptyset$. Therefore, the lines 5, 7, 13

Algorithm 5.4 Subchannel Allocation (SA) in the BS

1: if slot type is A, set $\mathcal{Q}' = \mathcal{Q}_0$, otherwise, set $\mathcal{Q}' = \mathcal{K}_0$
2: Initialize $q_k^0 = Q_k^0, r_{kn}^0 = B\log_2(1 + \frac{\hat{P}_0 e_{kn}^0}{|\mathcal{N}_0|\Gamma_k}), k \in \mathcal{Q}', n \in \mathcal{N}_0$.
3: **while** $\mathcal{N}_0 \neq \emptyset$ and $(\sum_{k\in\mathcal{Q}'} q_k^0 > 0)$
4: Compute $w_k^0 = q_k^0, k \in \mathcal{K}_0$
5: Compute $w_k^0 = (q_k^0 - Q_k^{m(k)}), k \in \mathcal{Q}' - \mathcal{K}_0$
6: if $Q_k^0 > BT\hat{s}, w_k^0 = w_k^0 + \rho_k^0 W_e, k \in \mathcal{Q}'$
7: if $L_k^{m(k)} - q_k^{m(k)} < BT\hat{s}, w_k^0 = -1, k \in \mathcal{Q}' - \mathcal{K}_0$
8: Compute $D_{kn}^0 = w_k^0 r_{kn}^0, k \in \mathcal{Q}', n \in \mathcal{N}_0$
9: Find $(k^*, n^*) = \arg\max_{k\in\mathcal{Q}', n\in\mathcal{N}_0} D_{kn}^0$
10: $x_{k^*n^*}^0 = 1$
11: $\mathcal{N}_0 = \mathcal{N}_0 - \{n^*\}$
12: $q_{k^*}^0 = \max(q_{k^*}^0 - Tr_{k^*n^*}^0, 0)$
13: if $k^* \in \mathcal{Q}' - \mathcal{K}_0$, then $q_{k^*}^{m(k^*)} = q_{k^*}^{m(k^*)} + \min(q_{k^*}^0, Tr_{k^*n^*}^0)$
14: **end while**

are not executed when SA algorithm is used by relays. Based on the above mentioned, the subchannel allocation task in EDDRA is split among the serving nodes, where the computational complexity of the SA algorithm in any node $m \in \mathcal{B}$ is of $O(|\mathcal{N}_m|^2|\mathcal{Q}_m|)$.

Step 4) Total Power Adjustment. After assigning the subchannels to the links, the BS and relays decide about the total power that they can distribute on their subchannels, to meet the constraints $C4, C7$. For this, based on SSP in (5.15), each node m solves the following problem, which we refer to as Total Power Adjustment (TPA), to find the total power, P_m, that it can use.

TPA (Total Power Adjustment Problems):

$$\max_{P_m} \sum_{n\in\mathcal{N}_m} \Big(BTw_{k(n)}^m \log_2(1 + \frac{P_m e_{k(n)n}^m}{|\mathcal{N}_m|\Gamma_{k(n)}})\Big) - IZ_m P_m, \forall m \in \mathcal{B} \tag{5.16a}$$

$$\text{s.t. } 0 \leq P_m \leq \hat{P}_m \tag{5.16b}$$

In the above, $k(n)$ indicates the index of the user, to the queue of which the subchannel n

has been allocated. The TPA problem is a convex problem with one variable; therefore, the optimal value, P_m^*, can be found easily by using an iterative one-dimensional search such as the Golden Section method, described in [70, Appendix C.3] , which has the computational complexity of $O(\log(1/\epsilon))$, where ϵ is the desired relative error bound.

Remark 3: As explained before, the constraint $C1$ is enforced over time through the virtual queues, $\{Z_m\}$, defined for that purpose. In fact, based on (5.7a), having nonzero Z_m means that in the past time slots, there have been the events of transmission with the total power, P_m, larger than the average power limit, P_m^{av}. Therefore, in TPA problem, Z_m applies a kind of negative feedback, to use less power than $\hat{P}_m$. The proposed importance factor I is in fact for amplifying this negative feedback to adjust the total power use in a short period of time. Without it, the second term in the objective (5.16a) would not be comparable to the first one, in a large period of time slots, before Z_m becomes big enough to impact the objective value. This is due to the fact that the values of the power variables are very small (in the order of 1-10 Watts), compared with the values of queue sizes multiplied by transmission rates (in the order of tens of Megabits).

After solving TPA problem, each node computes the power on its subchannels as follows, considering equal power distribution and noting that the rate on each subchannel can not be larger than $B\hat{s}$ (due to the limited spectral efficiency of modulation schemes in practice):

$$p_n^m = \min\left(\frac{P_m^*}{|\mathcal{N}_m|}, \frac{(2^{\hat{s}}-1)\Gamma_{k(n)}}{e_{k(n)n}^m}\right) \tag{5.17}$$

The reason for considering the term with $\hat{s}$ in (5.17) is to prevent using power more than needed for maintaining the desired bit error rate. It is obtained based on (5.1) and $C4$.

After the above steps, based on the variables x_{kn}^m, p_n^m, each node notifies its users about the subchannel allocations and the assigned transmission rates. Then, it transmits to them

and updates its actual and virtual queues.

5.3.5 Efficient Dynamic Centralized Resource Allocation

In this subsection, we briefly describe the Efficient Dynamic Centralized Resource Allocation (EDCRA) method, in which the BS performs all the procedures for resource allocation. In a centralized scheme, the BS needs to get notified about the channel states of all the links in the system, over all the subchannels[10]. For this purpose, since the indirect users do not have connection to the BS, the relays report to the BS about the channel conditions of the access links (which already the indirect users have reported to their serving relays). This imposes a signaling overhead of $O((K - |\mathcal{K}_0|)N)$ from relays to the BS. Considering the fact that in practice, the number of users is remarkably more than the number of the relays, the signaling overhead in EDCRA is a lot more compared with EDDRA[11] (which is of $O(\min(K - |\mathcal{K}_0|, N) + MN)$).

Having all the information about channel states and queue sizes, the BS performs STD procedure and if the slot type is set to A, it uses the SA algorithm as in EDDRA. However, if the slot type is set to B, the BS does not need to run SSD algorithm. Instead, it uses SA algorithm, considering all the subchannels and all the links, as follows. The queues in relays are assumed to be located in the BS and their corresponding access links are assumed as direct links starting from the BS to the indirect users; however, the weights and channel rates are considered the same as those of actual access links. Then, the SA algorithm is exploited to decide about the subchannel allocation to the different links in the system, which imposes the computational complexity of $O(N^2(\sum_{m\in\mathcal{B}}|\mathcal{K}_m|)) = O(KN^2)$. After

[10]In a centralized implementation, the BS has information about all the queue sizes, due to the fact that it has the history of the transmissions from all the queues.

[11]In the above discussions, we excluded the signaling overhead of channel state feedbacks from the receiver side of any link to the transmitter side, due to the fact that it is the same in EDDRA and EDCRA.

that, based on the corresponding subchannel allocations for all the nodes, i.e., $\{x_{kn}^m\}$, the BS specifies the powers to be used by each node by performing the total power adjustments for each node. Finally, the BS informs all the relays about the subchannels and powers they can use.

5.4 Performance Evaluation and Discussion

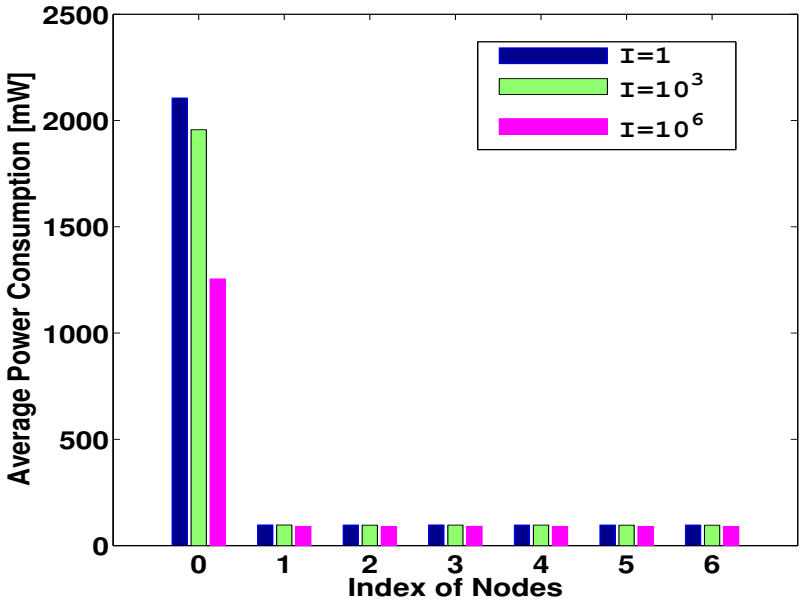
To evaluate the performance of the proposed algorithms, we have conducted extensive Matlab simulations for a system with 6 relays, which are located at the distance of 2/3 cell radius from the BS and in an equal angular distance from each other. The cell radius is 1000 m and the BS, relay and user antenna heights are considered 15 m, 10 m and 1.5 m respectively. Path loss attenuation is computed based on [61], the noise power spectral density in the receivers is -174 dBm/Hz and the users' BER requirements are 10^{-6}. The system bandwidth is divided into subchannels with the bandwidth of 180 kHz and the transmissions are done over the time slots of 1 ms. Data packets of 5 kbits arrive in the top layer buffers of the BS according to Poisson processes and if the corresponding buffer is not full, they are queued until getting admitted into the corresponding MAC layer buffers.

For the links between the BS/relay and users, Rayleigh channel model is used and for the links from the BS to relays, Rician channel model with κ factor equal to 6 dB [62]. Utility function is considered as $\mathcal{U}(a) = \log(a)$ for providing proportional fairness. Due to the large packet sizes which resulted in large queue sizes, based on the observations from simulation results, we have chosen V to be 10^7. This gives high value for utility function in (5.8a), to be comparable to the terms related to the weighted transmission rates. The buffer capacities at the BS and relays are considered equal to 100 and 10 packet sizes, respectively. The highest order for modulation is considered to be 64 QAM which has the spectral efficiency of 6 bits/sec/Hz.

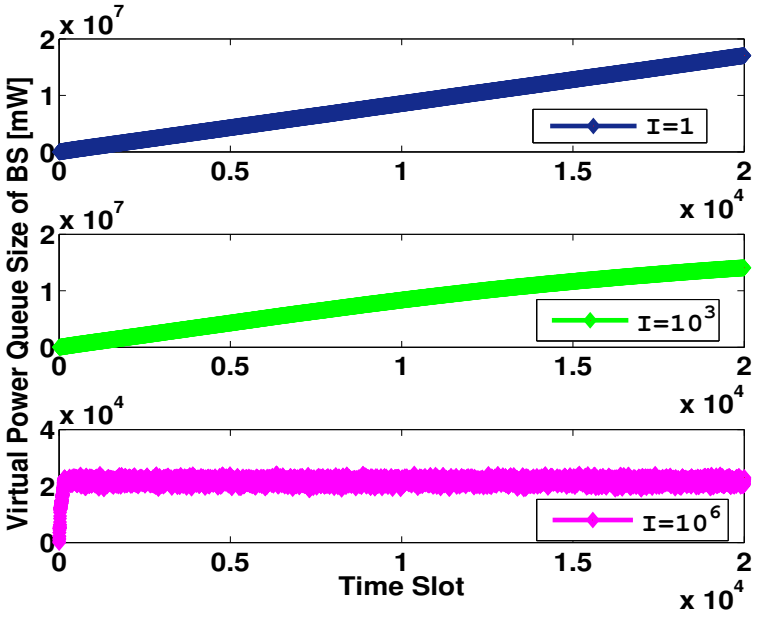
In the following, we first consider a special scenario to show the effect of parameter I. In this simulation, the number of the subchannels is considered equal to $N = 7$, the BS peak power equal to $\hat{P}_0$=34 dBm and the peak power of relays equal to $\hat{P}_m$=25 dBm, $m \in \mathcal{M}$. The average power constraint of the nodes are half of their peak power constraints, i.e., 31 dBm=1259 mW for the BS and 22 dBm=158 mW for the relays. There are 12 users in the system, 6 of them connected directly to the BS and the rest connected to relays, one user per relay. The distance of the direct users from the BS and indirect users from the corresponding relay is 300 m. The data arrival rate of each user is 100 packets/second, or equivalently 500 kbps.

Fig. 5.2(a) shows the average power used by each node, over 20000 time slots, with different values for importance factor I, and Fig. 5.2(b) depicts the virtual queue size corresponding to average power constraint of the BS, during the mentioned period. For choosing a suitable value for I, we have tuned it as stated in Section 5.3.1; i.e., in the range of the values for weighted rates in (5.8a) which is around 10^6 in our simulation settings. In the Figs. 5.2(a) and 5.2(b), we have also shown the case with $I = 1$ which is the case considered in the literature and is equivalent to ignoring the weight for the virtual power queues. Moreover, as a case between $I = 1$ and $I = 10^6$, we have shown the geometric average of them, i.e., $I = 10^3$, instead of arithmetic average, because the difference in the performance is more clear with the steps in the orders of the powers of 10.

It is observed that without considering a suitable I, the size of virtual queue in the BS grows constantly and the average power used over this period is about 2000 mW, far beyond the constraint of 1259 mW. This happens due to the fact that in equation (5.16a), the value of the first term is very large compared with the value of second one and as a result, it does not affect the optimization objective much; therefore, the only thing that limits the total power used is the peak power or maximum spectral efficiency. The consequence of



(a)



(b)

Figure 5.2: Effect of parameter I on (a) average power consumption of BS and relays (b) virtual power queue size of BS over time; $N = 7$, $|\mathcal{K}_0| = 6, |\mathcal{K}_m| = 1, m \in \mathcal{M}$.

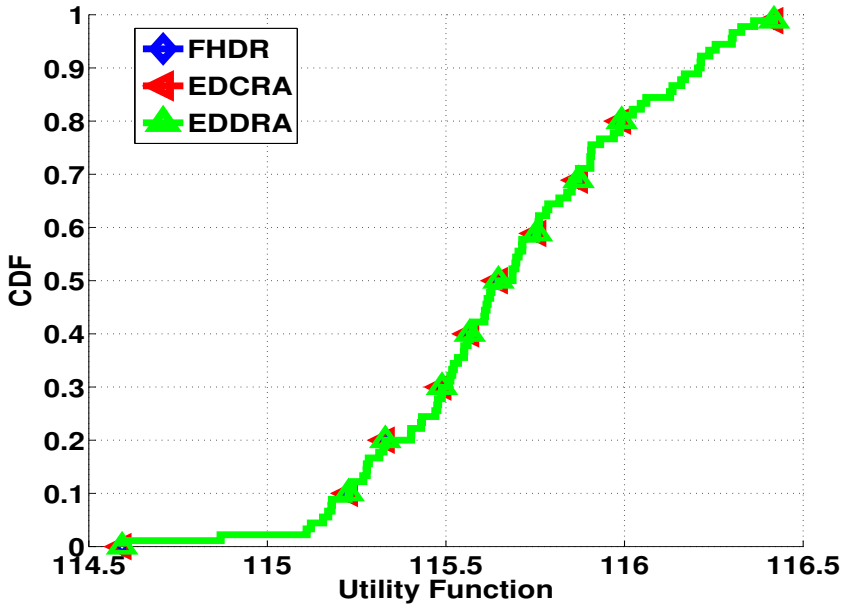
this is the steady use of the peak power of BS in each time slot, and the steady growth of its virtual power queue size according to equation (5.7). Without suitable I, this would continue for a long time, until the size of the virtual queue has grown so large that the first term in (5.16a) is comparable to the second one. However, by using a large I, this is prevented and the BS virtual power queue gets bounded after about 300 time slots, and the average power used in the whole simulation period is about the defined constraint. Note that due to fewer transmissions from the relays, compared with the BS, their virtual power queues did not grow large and remained stable in all the above values for I and had similar graphs as that of the BS in the case of $I = 10^6$. Due to this similarity, their graphs were omitted.

To investigate the overall performance of the proposed algorithms in general scenarios, we consider a system with 25 users, which are uniformly distributed in the cell area and are connected to the node from which they receive higher signal strength. The simulations are conducted for 100 runs, each over 10000 time slots, to generate different realizations of users locations. All the users have the data arrival rate of 20 packets/second or equivalently, 100 kbps, and the buffer capacity in the BS and relays are respectively 100 and 10 packets per user. There are 14 subchannels in the system, the BS peak power is 37 dBm, relays' peak power is 28 dBm and the average power constraints are half of the peak powers. As a benchmark, we have adapted the low-complexity centralized algorithm proposed in [48] to our system model, which we refer to as Fixed Half-Duplex Relaying (FHDR) in the figures. With FHDR, the odd numbered time slots are used for transmissions from the BS and the even numbered slots for transmissions only from the relays. The subchannel allocations in even numbered slots are based on considering a minimum of $\lfloor N/M \rfloor$ subchannels for each relay and assigning them based on Hungarian algorithm. For FHDR, the average power limit of each node is equally distributed over all subchannels, considering the maximum

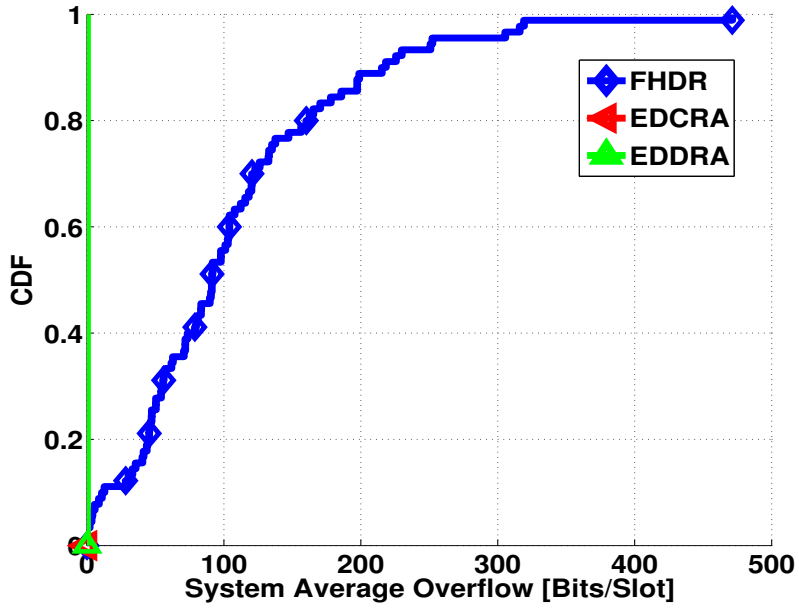
spectral efficiency constraint. Also, we have implemented the data traffic control procedure in the FHDR to compare the utility functions.

We note that due to limited buffer capacities, all the queues are stable and their sizes are less than buffer capacities. Therefore, in the following, we do not present any results about them and instead, we study the overflow performance. Fig. 5.3 displays the CDF of the system utility and total overflow from buffers of the relays. It is observed that even though all the algorithms have the same utility for data admissions, FHDR has the incidents of overflow. The result of this is lower throughput with FHDR, as shown in Fig. 5.4. On the other hand, the proposed centralized and distributed algorithms outperform FHDR and result in zero overflow and higher throughput. There are several reasons for this. EDCRA and EDDRA estimate power usage on subchannels and adjust it after the subchannel allocations. In subchannel allocation, they do not consider a minimum number for the nodes; instead, they allocate subchannels based on the higher demands and take into account the limited buffer capacity of the relays. All of the above lead to efficient use of the system resources and result in zero overflow in the buffers of relays, and higher throughput for the users.

Next, in the same system, we investigate the effect of increase in the data arrival rate on the performance of the proposed algorithms. Fig. 5.5 shows the system average utility and average total overflows. It is observed that as the data arrival rates increase, the utilities of the EDDRA and EDCRA also increase; but after the arrival rate of 60 packets/second, this increase is not much. This is firstly due to the fact that the utility function is a concave function which has diminishing returns as the arrival rates increase. The other reason is the fact that the system capacity is saturated in high packet arrival rates and the queues can not be served much and this prevents more admission of data into the system. Similar effect is observed about FHDR; however, since it does not determine the type of time slot



(a)



(b)

Figure 5.3: CDF of (a) system utility (b) total overflow from the buffers of relays; at the data arrival rate of 20 packets/second

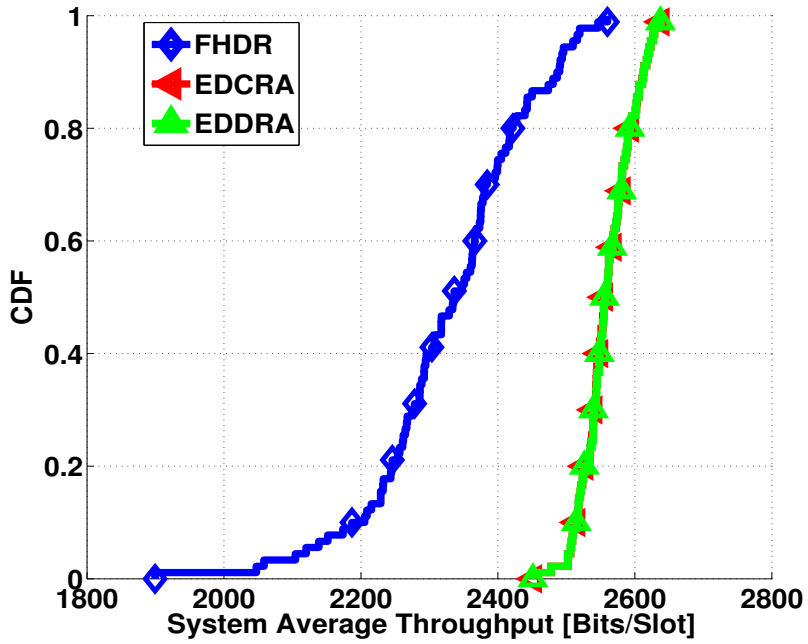
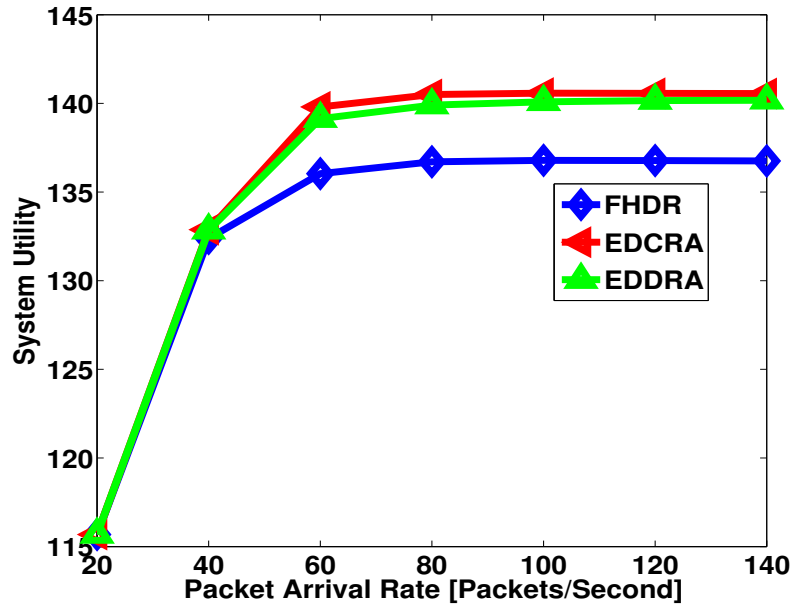


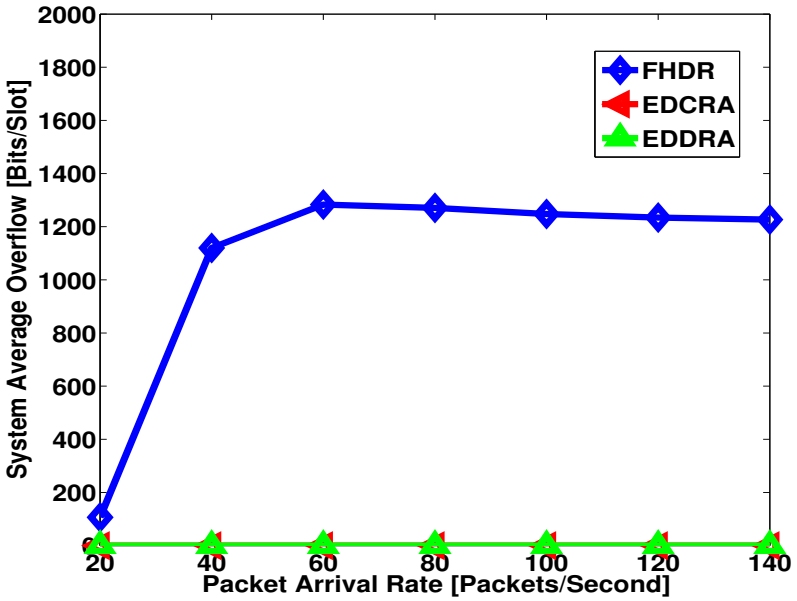
Figure 5.4: CDF of the system throughput at the data arrival rate of 20 packets/second

based on the demands and does not use the subchannel and power resources efficiently, it leads to higher queue sizes in the BS and consequently, lower admissions in high packet arrival rates. On the other hand, it does not take into account the buffer capacities of the relays and leads to data overflow. Due to the above mentioned, it results in lower throughput compared with the EDDRA and EDCRA, which is shown in Fig. 5.6.

It is also worth noting that the performance of the EDCRA is a little better than that of EDDRA. This is due to the fact that in EDCRA, the BS has information about the channel states of all the links and performs subchannel allocation based on individual demands of relays' and its own queues; but in EDDRA, the BS determines the subchannel sets of the nodes based on their *average* demands and then, each set of subchannels are only used to serve the set of the queues of the corresponding node. However, the degradation in the performance of EDDRA is not significant. The reason for this is the fact that there are usually several users for each node, with different channel conditions, that make it possible



(a)



(b)

Figure 5.5: Effect of increase in the packet arrival rate at the BS on (a) the system average utility (b) average overflow from the buffers of relays

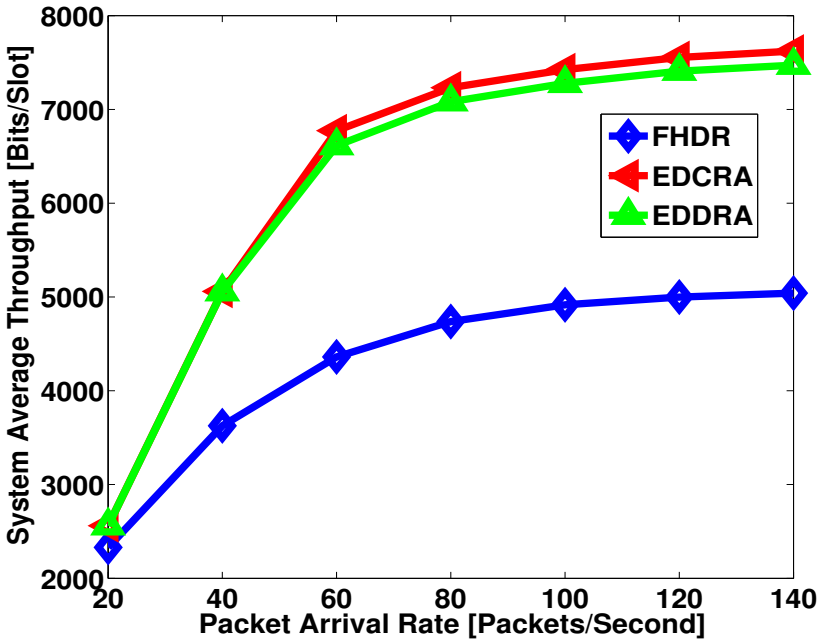


Figure 5.6: Effect of increase in the packet arrival rate at the BS on the system average throughput

for each node to utilize its set of subchannels efficiently. Also, considering an upper bound for the number of subchannels that each node can get prevents wasting of the resources. These observations show that using EDDRA, the system can allocate resources efficiently with less computations at the BS, low signaling overhead and without remarkable reduction in the system performance.

Next, in order to show the effectiveness of the proposed parameters ρ_k^m and W_e, we show data admission for direct and indirect users. It is observed in Fig. 5.7 that as the packet arrival rate of the users increases, the data admission for direct users increases while less data are admitted for indirect users. As explained before, this is due to the fact that the queues of indirect users in the BS get low weights and grow large which result in lower data admissions for them. To prevent that, ρ_k^m can be set to one and an extra weight of W_e can be used in the weights of the links from the BS to relays and the links from the relays

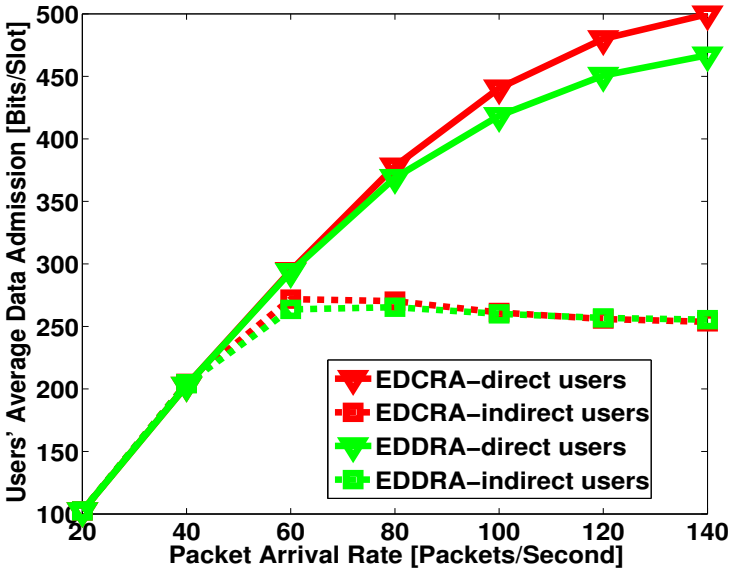


Figure 5.7: Effect of increase in the packet arrival rate on the amount of data admitted for direct and indirect users

to users; this will increase their chances for getting more service for the corresponding queues and consequently more data admissions. Fig. 5.8 shows this in the case of arrival rate equal to 140 packets/second for every user. It is observed that giving higher weights to the links from the BS to relays and the links from the relays to users can increase the data admissions for indirect users; this comes at the cost of large reductions in the data admissions of direct users. It is because adding extra weights results in the increase of subchannel allocations to the queues of indirect users in the BS and relays. Since the BS has higher power than the relays, the less subchannel allocation to the direct users means that their queues lose higher transmission rates than those for the queues in relays. As a result, the queues of direct users in the BS grow quicker and limit their data admissions.

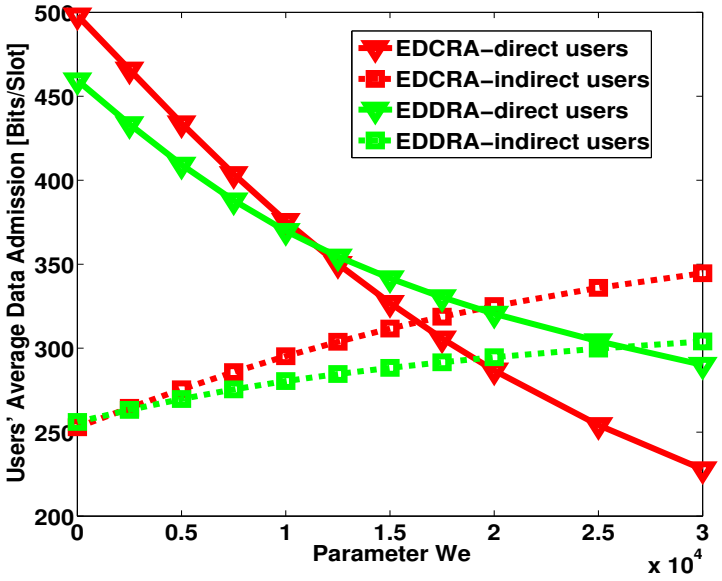


Figure 5.8: Effect of parameter W_e on providing fair data admission for direct and indirect users, at the arrival rate of 140 packets/second for every user

5.5 Summary

In this chapter, we have studied data admission control and resource allocation in buffer-aided relay-assisted OFDMA networks. We have formulated time slot, subchannel and power allocation as a utility-based stochastic optimization problem, taking into account several practical constraints. Using the Lyapunov drift-plus-penalty policy, we have transformed the problem into instantaneous subproblems. We have introduced important parameters that should be considered in using drift-plus-penalty policy in cellular networks. For practical considerations, we have proposed low-complexity strategy for allocation of power and subchannels and used it in distributed and centralized schemes. In particular, the proposed EDDRA policy is attractive for use in practice, due to its low-complexity as well as low signaling overhead. Numerical results confirm the effectiveness of the proposed parameters and also show that the proposed algorithms lead to significant improvement in data admission and throughput of the system.

Chapter 6

Efficient and Fair Throughput-Optimal Scheduling

6.1 Introduction

Wireless RSs are promising solutions for expanding cellular networks due to their low-complexity and cost effective deployment possibility [5, 72]. As a consequence of the advantages resulted by the exploitation of buffers in the RSs, it is expected that the buffer-aided relay-based cellular networks will attract a lot of attention in the coming years.

In the previous chapters, we addressed QoS provisioning and low-complexity issues in resource allocation in different scenarios of buffer-aided relay-assisted cellular networks. The studied instantaneous resource allocation problems were based on the MW policy to provide queue stability. In particular, the objective functions in those problems consisted the product terms of the transmission rates of the links and the differential backlogs of the corresponding queues at the two ends of the links. MW is well-known for being throughput-optimal in wireless networks with fixed number of queues and stationary ergodic data arrival processes [60, 73, 74]. It is also called Maximum Differential Backlog (MDB), which leads to backpressure routing method and is able to route packets to their destination without knowing the path to their destination and only based on the differential backlogs. Due to this, MDB is of great interest in the systems with multiple paths from the source

to the destination as in the system considered in Chapter 3 and the systems without infrastructure such as Ad Hoc networks.

However, in a topology with fixed routes and only one path for packet transmission from source to destination, as in the systems considered in Chapters 4 and 5, MDB leads to discrimination against the relayed users in terms of queueing delay. This happens because MDB gives lower weights to the links between the BS and RSs, which causes the relayed users' data to experience higher delay until reaching the user. Recently, [46] has proposed another throughput-optimal policy, called MSB, which tries to provide fairness in terms of packet delay, among the relayed and direct users. In MSB, the weight given to a transmission link from the BS is equal to the size of the corresponding queue in the BS, while the weight considered for a link from the RS is equal to the sum of the corresponding queue sizes in the BS and RS. However, MSB can lead to inefficient utilization of resources by granting channel to a queue in the RS which is almost empty; this can cause instability in some scenarios, because it leads to backlogs at the BS while allocating channel to the RS which does not have much data to transmit. Another drawback of MSB and MDB is that since the BS or RS need to have information about the queue states of each other, even a distributed system with independent channels for the BS and RS has the overhead of signaling about the queue states of the relayed users.

Motivated by [46, 60, 73, 74], in this chapter, we aim at designing new throughput-optimal scheduling rule in single-path relay networks, which is fair to the users served by the RSs and at the same time promises low signaling overhead. We propose MMW policy which has the following properties:

- Weights of the links are based on the local queue sizes either in the BS or RS, almost all the time. This makes them suitable for use in both centralized and completely distributed implementations, in the scenarios with shared or independent channels

for the BS and RS.

- Based on the above mentioned property, it is also possible to provide fairness in terms of queueing delay for direct and relayed users.
- It is possible to extend the algorithm for multihop cellular networks, without much extra signaling about the relayed users' channel and queue states, and, therefore, have completely distributed resource allocation.

The intuition behind MMW is based on the similarity between the topology of the single-path relay networks and that of parallel queueing system, which allows to adapt the MW algorithm to such networks and make it efficient in scheduling and resource allocation. In this chapter, we discuss MMW for the systems with single channels for transmissions from the BS and RS; however, the arguments presented can also be applied to multichannel systems.

The rest of this chapter is organized as follows. Section 6.2 describes the system model, provides a background on throughput-optimal scheduling and discusses the problem. In Section 6.3, we present the proposed method in detail and discuss its stability. Section 6.4 describes decentralized implementations, followed by simulation results in Section 6.5. Conclusions are stated in Section 6.6.

6.2 Preliminaries

6.2.1 System Model

As it is shown in Fig. 6.1, we study the downlink of a time slotted wireless network with a total of K wireless users in the system. Users are indexed by $k \in \mathcal{K}$, where $\mathcal{K}$ indicates the set of all the users. We consider a two-hop network with a single RS, as a basis for relay-

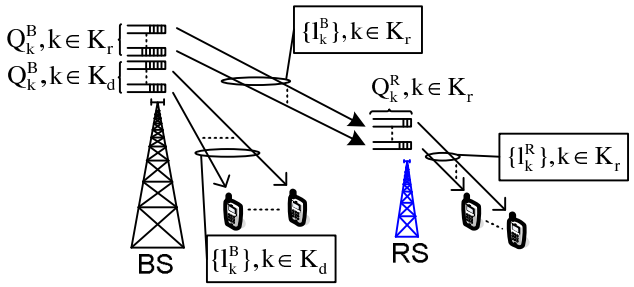


Figure 6.1: System model

based cellular networks. However, as will be discussed in Section 6.4, it is straightforward to generalize the algorithms and analysis for a system with several RSs and/or more than two hops. We assume that the set of the users $\mathcal{K}_d = \{1, 2, \cdots, K_1\}$ are served by the BS ("direct" users), and we use l^B_k to represent the link from the BS to user k. Users in the set $\mathcal{K}_r = \mathcal{K} - \mathcal{K}_d = \{K_1 + 1, \cdots, K\}$ have no direct link to the BS and are served through the RS ("indirect" users). For these users, l^R_k represents their link from the RS to user k and l^B_k represents their link from the BS to RS. Note that the links are specified by their beginning point and the user for which they carry data. Based on this, when $k \in \mathcal{K}_r$, l^B_k represents the link starting at the BS for transmission of data destined for user k. Therefore, this link ends at the RS, as there is no direct link between the BS and user $k \in \mathcal{K}_r$. This definition of notation is selected to remove the need for another superscript/subscript to specify the ending points of the links.

We consider two system scenarios. First, with a single wireless channel which is shared by the BS and RS and only one of the BS or RS can use the channel for transmission. Therefore, in each time slot, based on the scheduling algorithm, only one of the links, i.e., the links starting at the BS $\mathcal{L}^B = \{l^B_k\}$, $k \in \mathcal{K}$, or the links starting at the RS $\mathcal{L}^R = \{l^R_k\}$, $k \in \mathcal{K}_r$, can get the channel. We denote the set of all the links by $\mathcal{L} = \mathcal{L}^B \cup \mathcal{L}^R$ and the selected link for transmission in time slot t by $l^{sh*}(t)$, in which the superscript sh means "shared". In the second scenario, the BS and RS have independent and orthogonal

channels and, therefore, they can use them simultaneously for transmissions on one of their own links. Therefore, in this case, there will be two selected links, $l^{B*}(t) \in \mathcal{L}^B$ and $l^{R*}(t) \in \mathcal{L}^R$. Independent channels might be realized over frequency (e.g. separate frequency bands) or over time (e.g. using odd numbered time slots for transmissions from the BS and even numbered slots for transmissions from the RS). In the case of independent channels over frequency, it is assumed that the RS has the capability to receive (through the BS channel) and transmit (through the RS channel) at the same time.

Note that in the case of shared channel, a centralized or semi-distributed scheduling strategy can be implemented to prevent the simultaneous transmissions and interference; but in the case of independent channels, there is no need for centralized scheduling and a distributed scheduling is more reasonable. In Section 6.4, we will discuss these in more detail.

Note that the two scenarios discussed here can be used as the basis for resource allocation in multichannel systems, in the cases that all the subchannels are shared or partitioned between the BS and RS. For example in LTE-Advanced standards, Type 1 relays (non-transparent relays which are able to appear like an independent BS to their users) might have independent and separate bandwidth for the BS-RS and RS-user transmissions (Type 1a relays) or use shared bandwidths for them (Type 1b relays)[75]. Furthermore, in any of these types, the BS-user transmissions might be done in the same or different bandwidth than that used for the BS-RS transmissions.

We assume that the channel conditions of the links are stationary and ergodic and remain constant during transmission period. Instantaneous capacity of a link is a function of the channel state and the power that is allocated to it; e.g., for the link l_k^B, we have

$$c_k^B(t) = cf(p^B, s_k^B(t)) \tag{6.1}$$

where, $s_k^B(t)$ is the channel state of the link of user k from the BS in time slot t, p^B is the BS transmission power, $cf(.)$ is the capacity function and $c_k^B(t)$ indicates the number of bits possible to transmit on the link l_k^B at time slot t. Based on the scheduling algorithm, the transmitted bits per slot on the link of user k from the BS can be expressed as follows:

$$r_k^B(t) = c_k^B(t) i_k^B(t), \tag{6.2}$$

where $i_k^B(t)$ is an indicator variable which is 1 if the corresponding link, l_k^B, is selected for transmission in time slot t, and zero otherwise. The instantaneous capacity and transmission rate of a link of an indirect user k from the RS, $c_k^R(t)$ and $r_k^R(t)$, can be computed similarly based on s_k^R, p^R and i_k^R. We use $\underline{i} = [i_1^B, ..., i_K^B, i_{K_1+1}^R, ..., i_K^R]$ and Π to denote the scheduling vector and the set of all the feasible scheduling vectors, respectively. Based on the system settings, different constraints might apply on the scheduling vector and therefore, $\underline{i} \in \Pi$ is used to show this. In the case of shared channel, Π will be the set of all vectors with length $2K - K_1$, which have one entry equal to 1 and all the others equal to zero. In the case of independent channels, Π will be the set of all vectors with length $2K - K_1$, which have one of the first K entries and one of the last $K - K_1$ entries equal to 1 and all the others equal to zero.

We note that the channel state of the links of indirect users starting at the BS are the same and equal to s_R^B, the channel state between the BS and RS. In other words, $s_k^B = s_R^B$ for $k \in \mathcal{K}_r$. Because of this, we use $\underline{s} = [s_1^B, ..., s_{K_1}^B, s_R^B, s_{K_1+1}^R ..., s_K^R]$ and $\underline{r} = [r_1^B, ..., r_{K_1}^B, r_R^B, r_{K_1+1}^R ..., r_K^R]$, which have $K + 1$ entries, to denote the channel states vector and rates vector, respectively. It is assumed that the BS and RS have separate power sources and use fixed power for their transmissions. Therefore, based on (6.1) and (6.2), we show the transmission rates of the links as functions of their channel states and the indicator variables, and use $\underline{r}(\underline{i}, \underline{s})$ to indicate the specific transmission rate vector achievable in the

system channel state $\underline{s}$ by the scheduling vector $\underline{i}$.

It is assumed that the users' data traffic streams are inelastic, i.e., the destination does not send feedback to the source for flow control. Therefore, an admission control mechanism is used in the BS to ensure that the arrival rates can be supported by the network capacity. We assume that the BS has a buffer for each of the users, where data packets arrive according to an exogenous process and are queued until transmission to a direct user or to the RS. Similarly, the RS has a buffer for each of the indirect users which receives data packets from the BS and keeps them until transmission to an indirect user. Each queue in the BS (RS) corresponds to a unique link starting at the BS (RS). We assume that the buffer capacity of user k in the BS/RS is infinite, and denote the size of data queued in the buffer in time slot t by $Q_k^B(t)/Q_k^R(t)$, which is updated based on the following equations:

$$\begin{aligned} Q_k^B(t+1) &= \max\left(Q_k^B(t) - r_k^B(t), 0\right) + a_k^B(t), \quad k \in \mathcal{K} \\ Q_k^R(t+1) &= \max\left(Q_k^R(t) - r_k^R(t), 0\right) + a_k^R(t), \quad k \in \mathcal{K}_r \end{aligned} \tag{6.3}$$

where $a_k^B(t)$ is the number of bits arrived at the BS for user k in time slot t, according to an exogenous arrival process. $a_k^R(t)$ is the number of bits arrived at the RS, which would be equal to $a_k^R(t) = \min(Q_k^B(t), r_k^B(t))$. As (6.3) shows, it is assumed that the arrivals occur after transmissions in each time slot. In other words, up to $r_k^B(t)$ or $r_k^R(t)$ bits from the corresponding queue in the BS or RS will be transmitted first, and then, new arrivals will be accumulated. We assume that the packet arrivals at the BS are stationary and ergodic processes and have finite mean and variance. In the following, we will use the capacity of a link interchangeably with that of its corresponding user, e.g. $c_{l_k^B}(t)$ and $c_k^B(t)$ for the link l_k^B. Also, in order to uniquely specify the links, we will denote with b_l and k_l the beginning point and the user corresponding to the link $l \in \mathcal{L}$, respectively.

Note that the system model and the notations defined in this section can be used, with slight modification, for the two-hop networks with more than one relays and similar architecture (where each user is assigned to either the BS or a single relay and therefore, there is only one path for packets' travel from the BS to each user). In such networks, all the users that are connected to relays can be considered as the indirect users connected to only one RS and the queues in the relays can be thought of queues located in this specific RS. Then, the link definitions and their corresponding variables would be the same as above; the only difference would be the fact that now, the channel states of the indirect users' links between the BS and RS will not be the same and therefore, the vectors of channel states and rates should be defined as $\underline{s} = [s_1^B, ..., s_{K_1}^B, s_{K_1+1}^B, ..., s_K^B, s_{K_1+1}^R ..., s_K^R]$ and $\underline{r} = [r_1^B, ..., r_{K_1}^B, r_{K_1+1}^B, ..., r_K^B, r_{K_1+1}^R ..., r_K^R]$.

6.2.2 Background and Problem Statement

One of the important goals in queueing systems is the stability of queues, meaning that the queues should remain bounded. Any scheduling algorithm that is able to reach this goal in a stabilizable system is called throughput-optimal. This means that such an algorithm would maximize the throughput, as long as the exogenous packet arrival rates at the queues of the source nodes of the system are supportable by any other algorithm. Note that this approach is different from capacity maximization, discussed in the literature with the assumption of infinitely backlogged buffers at the source nodes. With that assumption, capacity maximization is equivalent to throughput maximization; however, when the queue dynamics are taken into account, capacity maximization does not necessarily lead to throughput maximization, due to burst nature of data traffic. Instead, queue stability is studied due to the fact that if an algorithm is able to make the system stable, in fact it will result in the maximum possible throughput which is equal to the sum of average rates

of the data arrived in the system.

In [73], authors addressed queue stability for the first time in a wireless multihop network. They proposed the MW algorithm which is throughput-optimal and is also known as MDB. It aims at maximizing the sum of weighted rates of the links, $\sum_{l\in\mathcal{L}} \beta_{k_l} w_l(t) r_l(t)$, where β_{k_l} is a constant related to the service class and priority of user k_l, and w_l is equal to the difference of the queue sizes of the nodes in the beginning and the ending points of the link l in time slot t. Therefore, MDB is able to utilize multiuser diversity and at the same time provides stability of the queues as long as all the packet arrival rates at the BS are strictly within the capacity region.

According to the MDB policy, in the case of a shared channel for the BS and RS, the channel is allocated to the "best" link based on the following criterion:

$$l^{sh*}(t) = \arg\max_{l\in\mathcal{L}} \beta_{k_l} w_l^{MDB}(t) c_l(t) \tag{6.4}$$

where

$$w_l^{MDB}(t) = \begin{cases} Q_k^B(t), & l \in \{l_k^B\}, k \in \mathcal{K}_d \\ Q_k^B(t) - Q_k^R(t), & l \in \{l_k^B\}, k \in \mathcal{K}_r \\ Q_k^R(t), & l \in \{l_k^R\}, k \in \mathcal{K}_r \end{cases} \tag{6.5}$$

In the case of independent channels for the BS and RS, (6.4) is used separately for each of the BS and RS channels, to allocate them to one of the links in $\mathcal{L}^B$ and $\mathcal{L}^R$, respectively.

MDB mostly has been used in the scenarios of multihop routing networks with mesh structure, where a packet can be forwarded through several paths to reach the destination. The fact of using the difference of queue backlogs as the weights of the links leads to the well-known "backpressure routing" method which aims at equalizing the differential backlogs, and routes the packets through a node that has small sizes of queues (which is a

sign of good channel conditions in its forward transmission links). In the case of single-path relay network, if all the users' queue sizes in the BS are equal, MDB will lead to lower priorities for the links $\{l_k^B\}$, $k \in \mathcal{K}_r$, when the indirect users have non empty queues in the RS. Therefore, on average, the total queue size of an indirect user (which is equal to the sum of its queue sizes in the BS and RS) will be different from that of a direct user with the same β_k, average data arrival rate and average channel conditions. As a result, MDB discriminates between the direct and indirect users by causing different queueing delays for them.

In order to provide a fairer solution in terms of delay, [46] proposed MSB algorithm for two-hop systems with independent transmission channels for the BS and RS. It uses a similar scheduling method as in (6.4) for each of the BS and RS channels, but with link weights, $w_l^{MSB}(t)$, different from those of MDB. MSB treats all the links starting at the BS, $\mathcal{L}^B$, similarly, and considers their weights equal to their corresponding queue sizes in the BS. On the other hand for transmissions from the RS to its users, the weights are equal to the sum of the corresponding queue sizes in the BS and RS; i.e.,

$$w_l^{MSB}(t) = \begin{cases} Q_k^B(t), & l \in \{l_k^B\}, k \in \mathcal{K} \\ Q_k^B(t) + Q_k^R(t), & l \in \{l_k^R\}, k \in \mathcal{K}_r \end{cases} \tag{6.6}$$

Although MSB aims at compensating the relaying effect for indirect users by using these weights, it might lead to inefficient utilization of system resources and consequently to instability, in the systems with shared channel for the BS and RS. This happens because an indirect user's link starting at the RS might get selected when its queue in the BS has large number of packets but its queue in the RS is almost empty.

Note that in the context of throughput-optimal algorithms and the networks with inelastic data traffic, throughput fairness is not discussed. This is due to the fact that

in this context, as a result of the stability of the system, each destination will receive an average throughput equal to the average data arrival rate in its source and can not get more or less than what has arrived for it. On the other hand, although throughput-optimal algorithms lead to bounded average queue sizes and average delays, they do not necessarily result in similar delay performance among the users. This needs to be addressed in cellular networks, because providing similar QoS, independent from the location of the users, is one of the objectives of the service providers. In order to discuss this rigorously, in this chapter we consider the scenarios where the users have equal data arrival rates, so the only difference between the users is their locations.

In the next section, we propose new throughput-optimal scheduling method for single-path relay networks, based on the MW algorithm and a new perspective.

6.3 MMW Policy

In this section, we propose the MMW scheduling policy, which can be exploited both in the systems with shared channels for the BS and RS and the systems with independent channels.

6.3.1 Motivation and The Main Idea

MMW aims at providing similar QoS among the direct and indirect users, efficient use of the system resources and lower signaling overhead, in single-path relay networks. It tries to maximize the sum of weighted rates of the links, with weights being different from those of MDB and MSB.

Fig. 6.2 depicts the main idea of MMW[12]. In the system of parallel queues as in

[12]Here, a single server indicates a shared channel. For independent channels, similar figure can be considered by an independent server for both queue 1 and 2 and an independent server for queue 3.

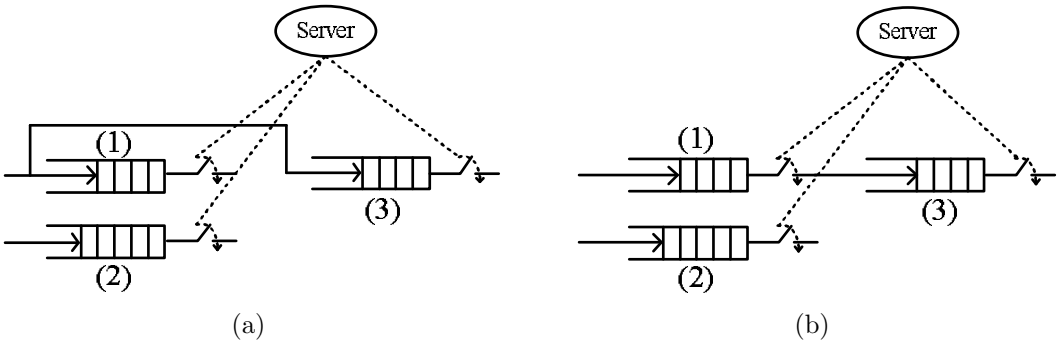


Figure 6.2: (a) Parallel queues (b) combination of parallel and tandem queues.

Fig. 6.2(a), all the queues have exogenous packet arrival processes. If such a system is stabilizable, MW algorithm will be able to provide stability by considering the weights of outgoing links equal to their queue sizes and maximizing the sum of weighted transmission rates of the links. Now consider Fig. 6.2(b), which is a combination of parallel and tandem queues, with the same channel capacities of the outgoing links and the same arrival processes in queues 1 and 2, as those in Fig. 6.2(a). Since the arrivals at queue 3 in Fig. 6.2(b) are delayed and processed versions of its arrivals in Fig. 6.2(a), i.e., with different packet sizes and arrival times, the total data entering queue 3 will not be more than those entering it in Fig. 6.2(a). Based on this, we propose to treat the system in Fig. 6.2(b) as parallel queues and, therefore, use the weight of each link proportional to the queue size in its beginning point. However, to be able to prove the stability, we consider a condition for that, which will be stated later.

Note that a single-path relay network is a system similar to Fig. 6.2(b), and therefore, we can exploit the perspective mentioned above. Specifically, we consider the weight of the link of an indirect user from the BS to RS, proportional to the corresponding queue size in the BS, unless the corresponding queue size in the RS exceeds a large threshold $\hat{Q}$; when this happens, the difference of the corresponding queue sizes in the BS and RS is used in its weight (similar to MDB). Also, to compensate the effect of relay's queueing delay on the

indirect users, we can use a coefficient based on the number of hops a packet experiences before arriving at the destination. For example, in a system with shared channel and two hops for indirect users, when all the links have similar channel conditions, this coefficient can be chosen equal to 2, to give twice as much priority for both links (starting at the BS and RS) of indirect users compared with the links of direct users. Furthermore, to improve the performance of proper link selection, we use actual channel rates of the links instead of their capacity.

Based on the above discussion, the objective in MMW is as follows:

$$\underset{\underline{i}\in\Pi}{\text{Maximize}} \sum_{l\in\mathcal{L}} \alpha_{k_l}\beta_{k_l} w_l^{MMW}(t) \min\{Q_{k_l}^{b_l}(t), r_l(\underline{i}(t), \underline{s}(t))\} \tag{6.7}$$

where β_{k_l} is the constant related to the service class and priority of user k_l, as before, and w_l^{MMW} refers to the weight of the link l in MMW, which is computed as

$$w_l^{MMW}(t) = \begin{cases} Q_k^B(t), & l \in \{l_k^B\}, k \in \mathcal{K}_d \\ Q_k^B(t) - I[Q_k^R(t) > \hat{Q}]Q_k^R(t), & l \in \{l_k^B\}, k \in \mathcal{K}_r \\ Q_k^R(t), & l \in \{l_k^R\}, k \in \mathcal{K}_r \end{cases} \tag{6.8}$$

and α_{k_l} is defined as

$$\alpha_{k_l} = \begin{cases} \frac{1}{c_{l,avg}}, & l \in \{l_k^B\}, k \in \mathcal{K}_d \\ \frac{z}{c_{l,avg}}, z \geq 1, & l \in \{l_k^B\} \cup \{l_k^R\}, k \in \mathcal{K}_r \end{cases},$$

In (6.8), $I[.]$ is the indicator function which is one if its argument is true, and zero otherwise. $c_{l,avg}$, is the estimated average transmission rate of link l, the use of which is clarified later.

Based on (6.8), as long as a queue size in the RS remains less than the threshold $\hat{Q}$, the weight of the corresponding link from the BS to RS would be proportional to just the queue

size in the BS. On the other hand, when the queue size in the RS gets larger than $\hat{Q}$, the weight of the corresponding link from the BS to RS will be proportional to the difference of the queue sizes in the BS and RS. As will be shown later through simulations, by defining a suitably large threshold, the queue sizes in the RS will be less than the threshold, almost all the time (more than 99.99 percent); therefore, practically, the weights of all the links will be proportional to just the queue sizes in their beginning point (similar to the system of parallel queues), i.e., $w_l^{MMW}(t) = Q_{k_l}^{b_l}(t)$.

In the definition of α, the parameter z is the coefficient mentioned before: if we want to compensate for the effect of relaying delay in a single-path multihop network, it can be adjusted to a value larger than one, based on the number of hops, channels and other network parameters. On the other hand, the reason for using $c_{l,avg}$ here is independent from the fact of different hops for the users and is similar to that of PF scheduling [76]. The inclusion of it in α_{k_l} is to compensate for the effect of different path losses of the links on the queue sizes. For example, consider two direct users with the same packet arrival rate and different distances from the BS. Although a throughput-optimal algorithm will keep their queues stable and provide the same throughput for them in long term, without using $\frac{1}{c_{l,avg}}$ the user far from the BS will have a larger queue size on average over time and, therefore, will receive its data with more delay. The min operator, used in (6.7), makes sure that the actual channel rates of links get considered. This has a similar effect as in single-hop networks [77], and will prevent the inefficient link selection (which might happen when a queue has not much data but its corresponding link has high channel capacity) and the consequent large queue sizes in the BS and RS.

Note that in the single-path relay networks, where the routes are fixed, there is no need for backpressure routing. Hence, defining a suitable threshold, such that its value is significantly larger than the RS queue sizes (as will be discussed in Section 6.5), makes it

possible to use the local queue sizes in the weights of the links. This way, MMW applies a kind of forepressure to the next hop and enforces the RS queues to remain below the threshold. Moreover, defining α_{k_l} as stated above, helps to provide more service to the indirect users' queues in the BS and RS, and improves the fairness in terms of queueing delay between the direct and indirect users.

Based on the above, in the case of a shared channel, MMW will allocate the channel to the link l^{sh*} as follows:

$$l^{sh*}(t) = \arg\max_{l\in\mathcal{L}} \overbrace{\alpha_{k_l}\beta_{k_l} w_l^{MMW}(t)\min\{Q_{k_l}^{b_l}(t), c_l(t)\}}^{V_l(t)} \tag{6.9}$$

In the case of independent channels for the BS and RS, (6.9) can be used separately for selecting l^{B*} and l^{R*} among $\mathcal{L}^B$ and $\mathcal{L}^R$, respectively. We will show later that MMW leads to an interesting result in this case: By selecting a suitably large $\hat{Q}$, there is no need for any information exchange about the channel states and the queue sizes of indirect users, almost all the time. Hence, a completely distributed stabilizing resource allocation can be implemented, which is highly scalable for multihop networks with more than two hops.

6.3.2 Stability Analysis

We note that according to [60], the capacity region of our system is defined as follows:

Definition 6.1 *The network capacity region Λ is the closure of the set of all rate vectors $\underline{\lambda}^B = [\lambda_1^B ... \lambda_K^B]$ that can be stably supported over the network, considering all possible algorithms (possibly those with full knowledge of future events).*

It is shown in [60] that the capacity region is determined based on the channel states, power allocations and achievable transmission rates for the links of the network. The sufficient condition for the stability of the network is to have $\underline{\lambda}^B$ interior to Λ, i.e., there should

exist an $\epsilon > 0$ such that $\underline{\lambda}^B + \epsilon \in \Lambda$ [60]. This ensures the feasibility to provide the system queues with the service rates strictly larger than their data arrival rates. Considering the above mentioned, in the following theorem, we show that MMW is throughput-optimal.

Theorem 6.1 If there exists an $\epsilon > 0$ such that $\underline{\lambda}^B + \epsilon \in \Lambda$, then, MMW is able to stabilize the system.

The proof of Theorem 6.1 is given in Appendix C.

6.4 Distributed and Semi-Distributed Implementations

The MMW scheduling algorithm can be implemented easily in decentralized ways. According to (6.5) and (6.6), in MDB (MSB), the BS (RS) needs the information about the queue sizes in the RS (BS) to compute its weights. On the other hand, according to (6.8), by defining the queue threshold large enough in MMW, the BS and RS can decide about their weights independently and based on only their local queue information, almost all the time. In the following, we assume that a suitably large $\hat{Q}$ has been selected (this is clarified in Section 6.5) such that the RS queue sizes remain below it, and based on that, we discuss the decentralized resource allocation. In the rare events that any RS queue size exceeds the threshold, the RS will need to notify the BS about the queue size, in addition to any other information.

6.4.1 Case1: Shared Channel

In the case of a shared channel, the BS and RS can initially exchange messages about the network parameters to decide the value of z. After that, using MMW in each time slot

and based on their local QCSI, they can compute V_l in (6.9) for their own links and the RS can inform the BS about its maximum V_l. Then, the BS can compare it with V_l of its own links and find l^{sh*}; If $l^{sh*} \in \mathcal{L}^B$, the BS will inform the ending point of l^{sh*} (which is either one of the direct users or RS) to prepare for data reception and then, the BS will use the channel to serve the corresponding queue and to transmit data to the link's ending point. If $l^{sh*} \in \mathcal{L}^R$, the RS can perform the similar procedures to transmit data, after getting notified by the BS that the winner link belongs to it[13]. Note that in the semi-distributed resource allocation here, the RS just needs to inform the BS about the highest value of its V_l for the transmission channel; but in the case of using MDB or MSB, it will also be needed to exchange QSI for the modified queue sizes in the previous time slot. Specially, this is of great importance in OFDMA networks, where usually there are a large number of subcarriers and users in the system and, therefore, in each time slot, it is possible to transmit to many users and have modified information about many queues. Specifically, considering a system with N^{sc} subchannels, with MDB this can impose an overhead of up to $\min(N^{sc}, K - K_1)$ signaling from the RS to the BS, as it is possible to transmit to this number of users from the RS; with MSB an overhead of up to $K - K_1$ signaling from the BS to the RS will be imposed due to the possibility of packet arrivals in any of the BS queues that belong to indirect users.

6.4.2 Case2: Independent Channels and Highly Scalable Framework

In the case of independent transmission channels for the BS and RS, distributed resource allocation becomes a lot easier and leads to a more interesting result. After initial system settings and adjusting $z > 1$, each of the BS and RS can use (6.9) for allocating its channel

[13]In the systems with more than one RS, the similar procedures can be performed: all the RSs inform the BS about their maximum V_l and then, the BS finds the l^{sh*} and notifies the RS to which l^{sh*} belongs.

and power to a link in $\mathcal{L}^B$ and $\mathcal{L}^R$, respectively. The main interesting feature here is that, in terms of signaling, the RS acts like a direct user and feeds back just its own receiving channel condition to the BS, and there is no need for any other signaling between the BS and RS, neither about V_l nor about the QCSI of indirect users[14]. Note that in this case, there will be no point in using centralized resource allocation, as the BS and RS have separate power resources and can operate independently. Accordingly, in OFDMA systems with independent frequency bands for the BS and RS, each of the serving nodes can independently use the algorithms proposed in the literature for traditional OFDMA systems (without relays) [68, 78–86] to allocate its resources.

Based on the above mentioned, MMW can be easily extended to single-path relay networks with more than two hops, as in Fig. 6.3, and independent channels for the BS and RSs, where the BS and RSs can use (6.9) for allocating their channels to their own links. This way, the aforementioned feature still holds and each RS acts as a direct user for the previous hop and as a BS for the next hop. Consequently, MMW provides a framework for completely distributed resource allocation in relay-based cellular networks with independent channels, which is highly scalable without increasing the signaling overhead between the serving nodes (i.e., the BS and RSs); as a result, it can be of great advantage in building relay-based cellular networks.

6.5 Performance Evaluation

To evaluate the system performance, we have conducted extensive Matlab simulations over 10000 time slots, in a system with cell radius equal to 1000 m, where the RS is located at the distance of 2/3 cell radius from the BS; the BS, RS and user antenna heights are

[14]With MDB and MSB, again an overhead will be imposed in the orders mentioned in the previous subsection.

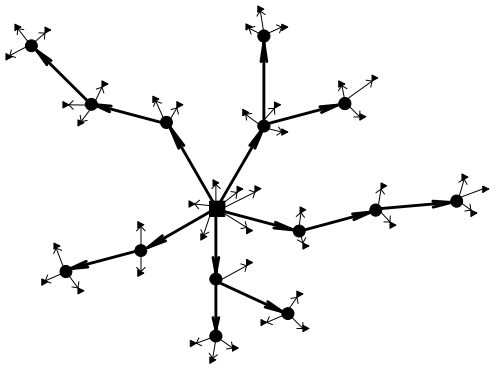


Figure 6.3: Multihop relay-based cellular network. Square, circle and triangle represent the BS, RS and user, respectively.

assumed 15 m, 5 m and 1.5 m respectively, and the transmission power for the BS and RS is respectively 31 dBm and 22 dBm. The path loss of each link is calculated based on [61] and the power spectral density of the noise in the receivers is assumed equal to -174 dBm/Hz. The bandwidth of the shared channel used for the transmissions from the BS or RS is considered equal to 1 MHz and the time slot duration is set to 1 ms. Users' data packet arrivals at the BS are based on Bernoulli distribution with packet sizes equal to 1 kbits.

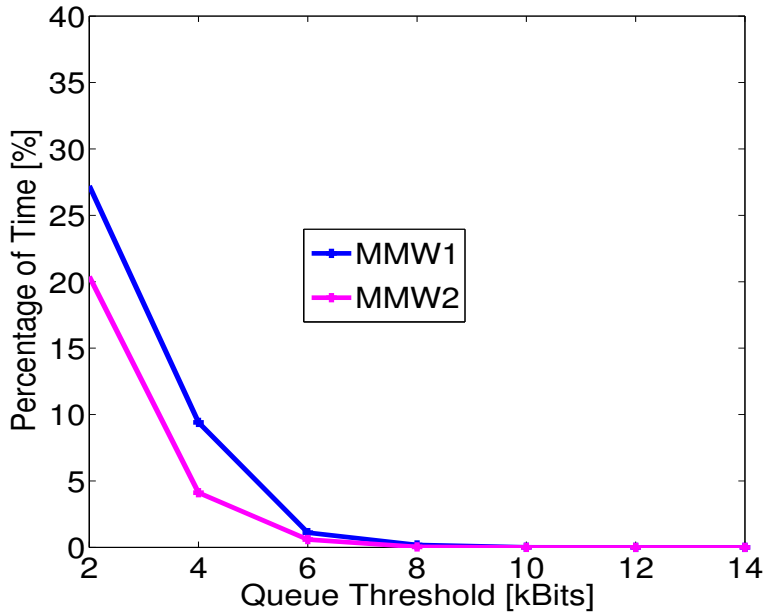
For the channel from the BS or RS to the users, Rayleigh model is used while the channel from the BS to RS is modeled as Rician with κ factor equal to 6 dB [62]. The channel fading over the system bandwidth is assumed to be flat which varies independently from one time slot to another.

For computing the channel capacity of the links, we have used $c_l(t) = WT\log_2(1 + p_l s_l(t))$, where W is the channel bandwidth, T is the time slot duration and p_l is the power that is used for transmission on the channel for link l. $s_l(t)$ is the instantaneous channel gain-to-noise ratio in the receiver side of the link l, which is computed as $s_l(t) = \frac{|h_l(t)|^2 G_l}{\sigma_n^2}$, where $h_l(t)$ and G_l are respectively, the small scale fading coefficient in time slot t and the path loss attenuation of link l, and σ_n^2 is the variance of Gaussian noise in the receiver

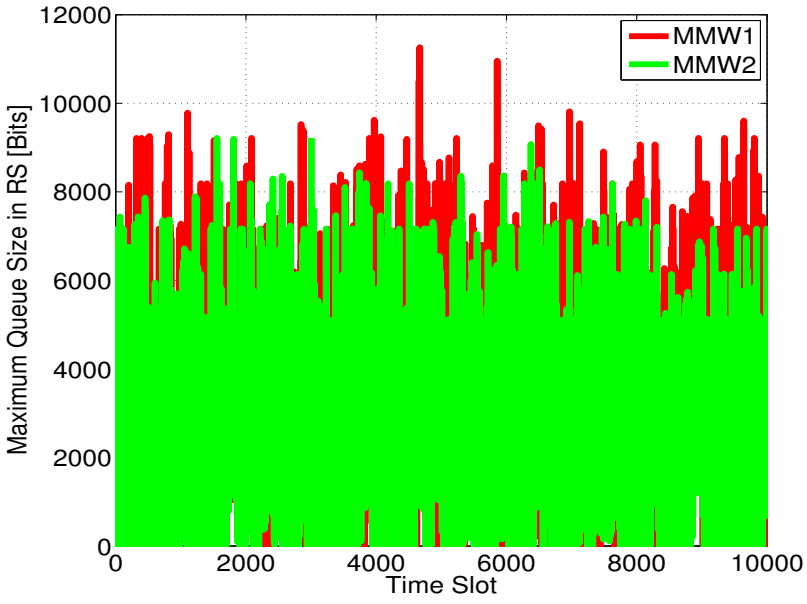
side of link l. Based on these, as an estimate of average channel capacities, we have used $c_{l,avg} = WT \log_2(1 + \frac{p_l G_l}{\sigma_n^2})$, i.e., the capacity without considering the small scale fading. In order to be able to investigate just the effect of distance and relaying on user delays, we have assumed all the users have the same service class and therefore, we have set $\beta_k = 1$ for all of them.

First, in order to decide about the queue threshold, we consider a system with a shared channel for the BS and RS in a scenario where 8 direct and 4 indirect users are located in the middle point between the BS and RS, and the middle point between RS and the cell edge, respectively. The reason for selection of these distances is to approximate the average case of random distances of direct and indirect users from the BS and RS, respectively. The packet arrival probability in each slot is 0.28 for every user. Therefore, the average bit arrival rates are equal to 280 kbps, which correspond to a medium load in this scenario. Fig. 6.4(a) displays the percentage of time slots in MMW[15] with $z = 1, 2$, that on average, each queue in the RS exceeds the shown preset queue threshold (which makes it necessary to set the weight of corresponding links from the BS to RS based on the differential queue sizes). It is observed that as the considered threshold increases, the percentage of time slots decreases and after the threshold of 12 kbits, this percentage gets zero. It is worth mentioning that in this setting, due to the forepressure effect stated before, queue sizes remain under the threshold. This can be observed in Fig. 6.4(b), which depicts maximum queue sizes in the RS when $\hat{Q} = 14$ kbits. These results show that if we preset the threshold to a large value, MMW will consider the weights of the links proportional to just the queue sizes in the links' beginning points, almost all the time. This will lead to less overhead which can facilitate the semi-distributed and completely distributed resource allocation, as discussed in Section 6.4. Also, it will improve the delay fairness in the system, as shown in the figures later. In the following, we have considered the queue threshold equal to 50

[15]In the figures, for brevity, the coefficients have been displayed immediately after the term MMW.



(a)



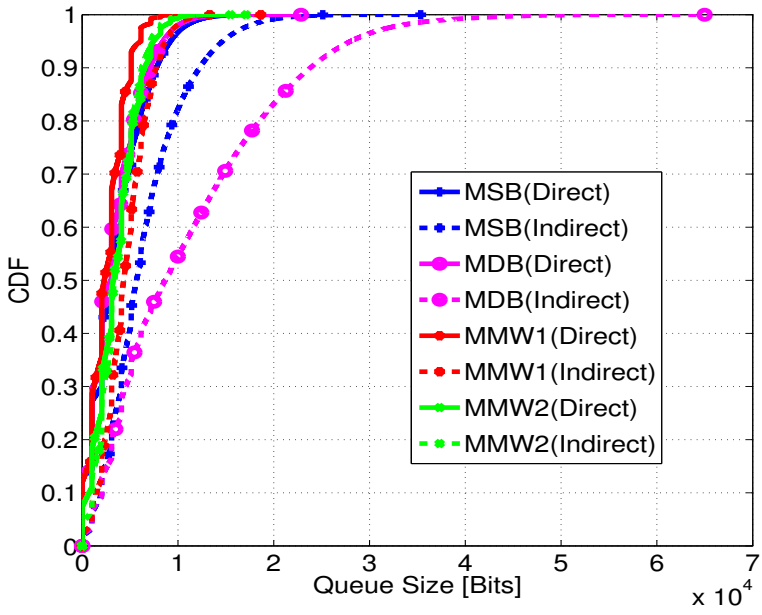
(b)

Figure 6.4: (a) Percentage of time slots each queue in the RS exceeds the queue threshold (b) maximum queue sizes in the RS, over time, when $\hat{Q} = 14$ kbits.

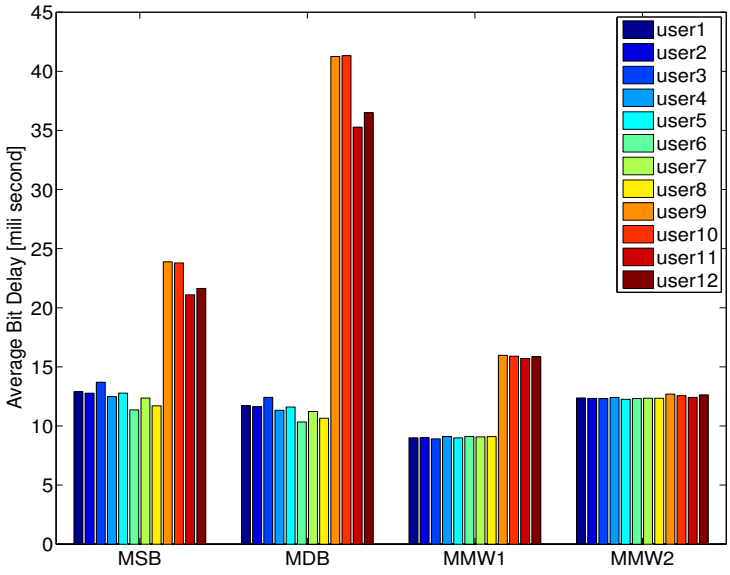
kbits and observed that almost all the time (more than 99.99 percent of the time slots), the RS queues remain under the threshold and therefore, local queue sizes are used in the weights of the links.

Now, we investigate the delay fairness in the above system, i.e., with a shared channel and packet arrival probability of 0.28, where the 8 direct users in the distance between the BS and RS and 4 indirect users between the RS and the cell edge are randomly located (with uniform distribution). We have run simulations for more than 100 realizations of user locations, each over 10000 time slots. Fig. 6.5 depicts the CDF of direct and indirect users' experienced queue sizes and their average bit delays. Here, the queue size for relayed users is the sum of their corresponding queue sizes in the BS and RS, and the average bit delay is defined based on the Little's law, as the ratio of the average queue size to the average data arrival rate [87]. A fairer policy will have closer values of queue sizes and average bit delays, and, as a result, closer graphs for both groups of users. It is observed that in a shared channel scenario, on average, MDB has good performance for direct users while discriminating against indirect users. This is due to the differential backlog based weights for the links of indirect users from the BS to RS, according to (6.5), which results in lower priorities and service rates for the queues of indirect users in the BS. Moreover, the queueing at the RS adds to the delay of relayed users' data before reaching the users. Therefore, on average, the indirect users receive their data with more delay than the direct users. MSB has better performance than MDB, as it assigns the weights based on (6.6) which results in higher priorities for the links of indirect users starting at the BS and RS, compared with MDB, and more service for their corresponding queues.

MMW1 and MMW2 behave better than MDB and MSB, in providing good delay performance for both the direct and indirect users. In particular, using $\min\{Q_{k_l}^{b_l}(t), r_l(i_l(t), s_l(t))\}$ in (6.7) leads to efficient link selection and utilizing the channel, and using $1/c_{l,avg}$ helps in



(a)



(b)

Figure 6.5: (a) CDF of the average queue size for direct and indirect users (b) average bit delay of all the users; the case of shared channel.

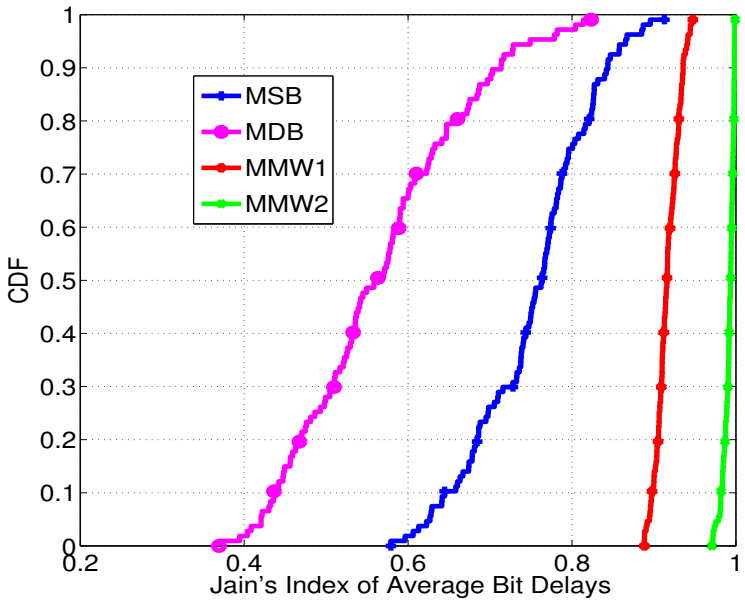


Figure 6.6: CDF of Jain's fairness index for average bit delays of the users; the case of shared channel.

providing similar delay for the users in each group. Moreover, by using $w_l^{MMW}(t) = Q_{k_l}^{b_l}(t)$ almost all the time, MMW1 and MMW2 provide more service for the queues of indirect users at the BS, compared with MDB. This reduces the delay for the data of indirect users. Furthermore, MMW2 sets $z = 2$ which doubles the weights of indirect users starting at the BS and RS, and serves their corresponding queues with even higher rates. This way, MMW2 decrease the delay of the indirect users and increases that of the direct users and, therefore, provides similar performance for all the users. These results are also reflected in Fig. 6.6, which shows the CDF of the Jain's fairness index [88] for users' average bit delays in each realization of user locations. It is observed that the Jain's fairness index for average bit delays in MSB is larger than MDB. MMW1 leads to even larger values of Jian's fairness index and MMW2 has the largest. Note that the Jain's fairness index does not provide any insight about the efficiency; it is by the figures of queue size and average bit delay that we get a comprehensive idea about the efficiency and fairness of the algorithms.

To have a clearer picture about the effectiveness of the proposed algorithms, we also present the performance in a higher load with data arrival rate of 350 kbps for each user and a special instance of user locations: all the direct users are located in the closest distance to the BS (50 m) and all the indirect users are on the cell edge, i.e., at about 333 m from the RS. As Figs. 6.7 and 6.8 show, in this scenario, all the algorithms lead to almost similar queue sizes and average bit delays for direct users, but MDB results in high values for indirect users. MSB is better than MDB, but the performance of MMW is remarkably better. Based on the above, we note that MMW is very suitable for giving priority to indirect users and has a fair performance. Specifically, MMW1 has a better fairness compared with MSB and MDB, and MMW2 is the fairest.

As it was explained in subsection 6.2.2, it should be noted that MSB was proposed

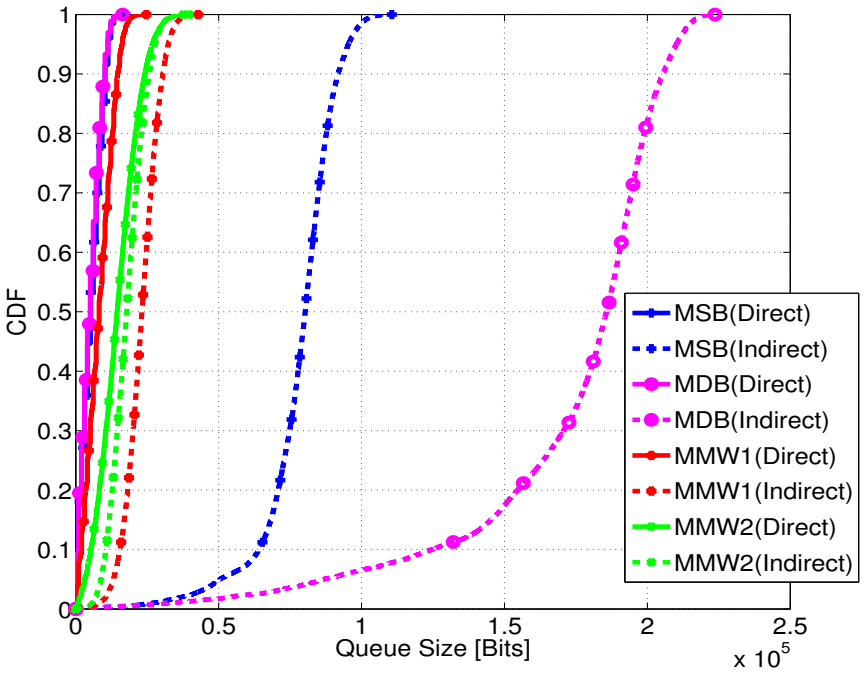


Figure 6.7: CDF of the average queue size for direct and indirect users; the case of shared channel, where direct users are located close to the BS and indirect users on the cell edge.

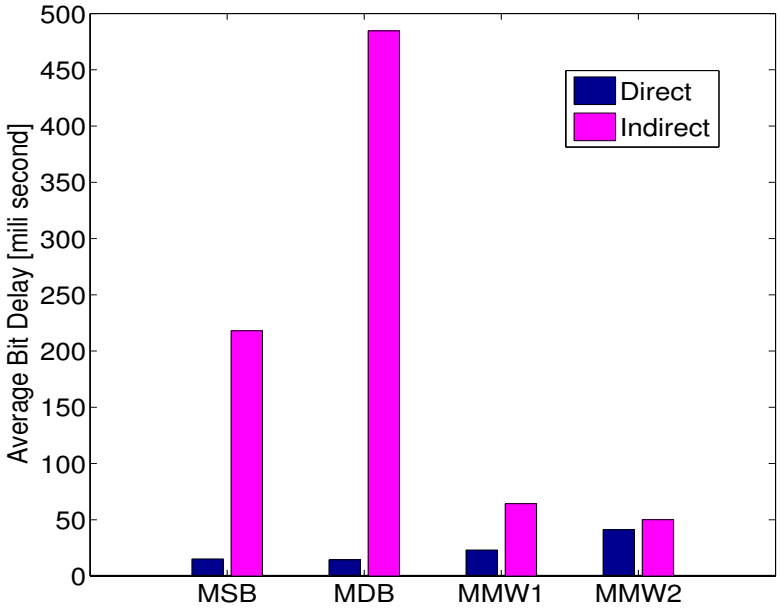
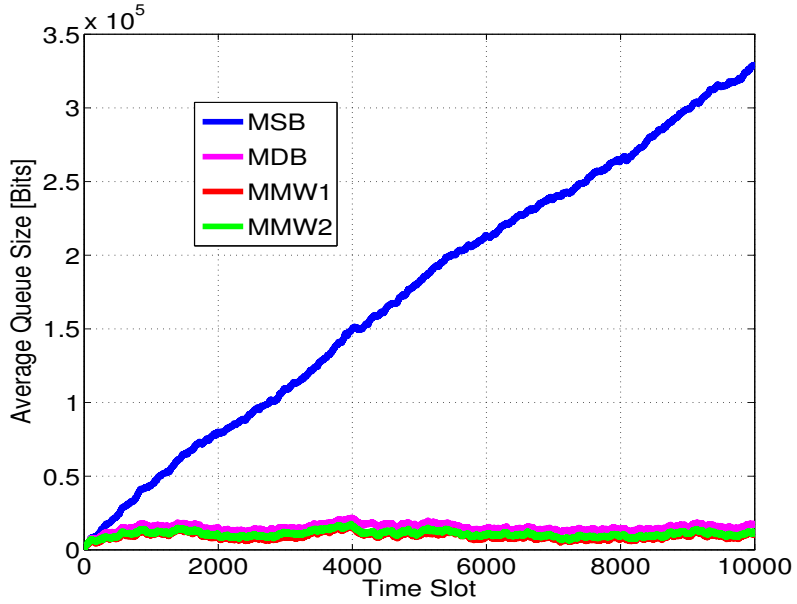


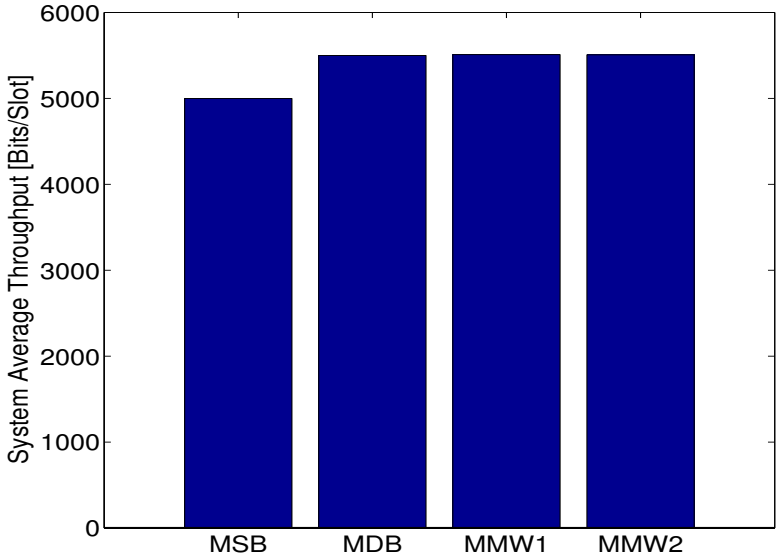
Figure 6.8: Average bit delay for direct and indirect users; the case of shared channel, where direct users are located close to the BS and indirect users on the cell edge.

in [46] for the case of independent channels for the BS and RS and it might lead to instability in some scenarios in the case of shared channel. To see this clearly, consider a scenario in which the relayed users are located close to the RS and at the distance of 50m from it. On the other hand, direct users are located in the middle point between the BS and RS distance. The packet arrival probability in each slot is 0.45 for every user, leading to the average bit arrival rates of 450kbps. Figure 6.9 shows the average queue size and average throughput of the system. It is observed that the queue sizes grow unbounded with MSB, whereas MMW and MDB are able to keep the queues stable. As a result, while with MMW and MDB, the system average throughput in each time slot is equal to the total average data arrival rate in the system (i.e., 12*.45= 5.4 kbits per time slot), MSB is not able to support the arrival rates.

As explained previously, this happens because an indirect user's link from the RS with



(a)



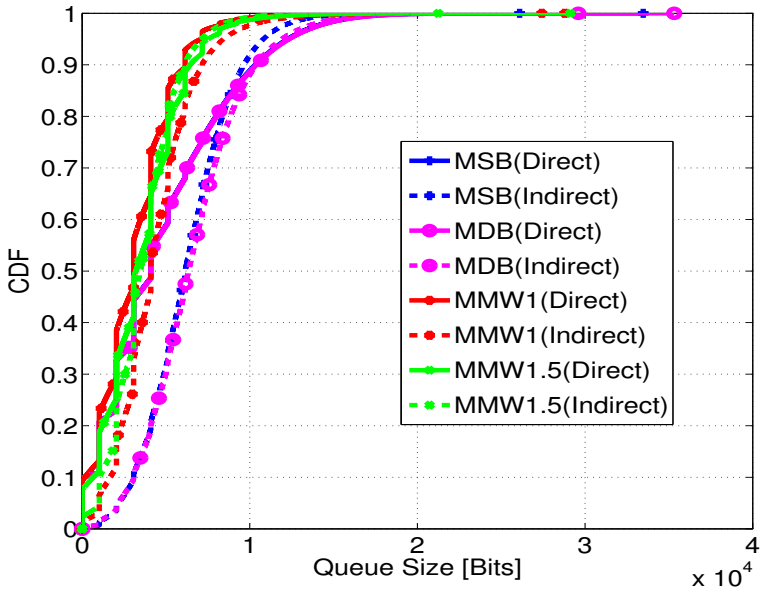
(b)

Figure 6.9: (a) System average queue size over time (b) system average throughput in each time slot; the case of shared channel, where indirect users are located close to the RS.

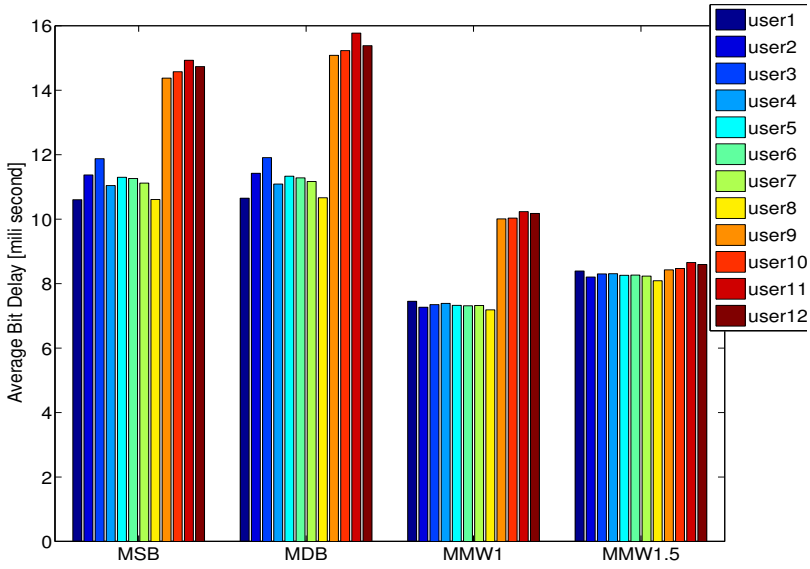
few bits in its corresponding queue in the RS will get selected when there are many bits in its corresponding queue in the BS. This happens specially in the cases that the average channel gain of the indirect users' links starting at the RS is higher than those starting at the BS. In this situation, MSB serves the indirect users' queues in the BS with a lower rate, and due to this inefficiency, its stability region is smaller.

For the case with independent channels, we have assumed a separate channel in the RS with the bandwidth of 500 kHz and over a different frequency band. Thus, the BS and RS can transmit at the same time without interfering to each other. Fig. 6.10 investigates the CDF of direct and indirect users' experienced queue sizes and the average bit delays. Users' location settings are similar to those of Fig. 6.5 and the packet arrival probability is equal to 0.42, which is a medium load in the case of independent channels. This is due to the fact that with separate resources for serving indirect users' queues in the RS, the system capacity is higher than the system with shared channel. It is observed that while MSB has better performance than MDB in some realizations, MMW behaves better than MDB and MSB, in all the realizations of user locations and result in lower queue sizes and average delays. In particular, MMW with $z = 1.5$ provides similar queue sizes and average delays for direct and indirect users and therefore, as shown in Fig. 6.11, has higher values for Jain's fairness index. The coefficient $z = 1.5$ has been obtained through simulations. Note that since the indirect users have the benefit of using two channels, one for their BS-to-RS links and one for their links starting at the RS, $z = 2$ will not be suitable coefficient if we want to have high fairness, as it can lead to discrimination against the direct users.

Similar to the case of shared channel, we also note that in the case of independent channels, MMW is more efficient as it assigns the links' weights, almost all the time, proportional to just the queue sizes in the beginning points of the links and, therefore, does not need the BS/RS signaling about the QCSI of indirect users.



(a)



(b)

Figure 6.10: (a) CDF of the average queue size for direct and indirect users (b) average bit delay of all the users; the case of independent channels.

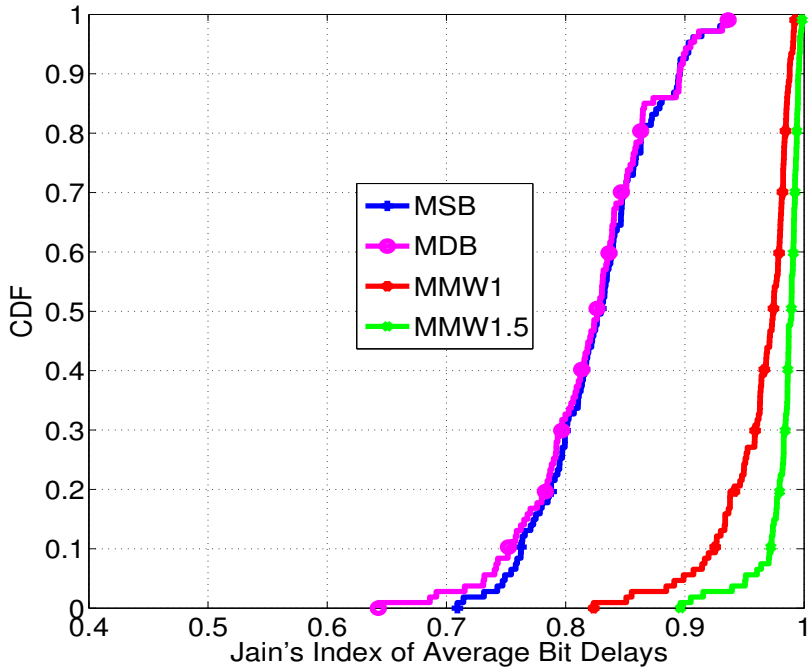


Figure 6.11: CDF of Jain's fairness index for average bit delays of the users; the case of independent channels.

6.6 Summary

In this chapter, we have proposed a variation of throughput-optimal resource allocation algorithms, namely MMW, for buffer-aided relay-based cellular networks with one path for packet transmissions from the BS to each user. MMW uses just the corresponding local queue size for assigning the weight of a link, unless in the rare events that the queue size in the RS exceeds a predefined large threshold, in which case the weight is defined according to conventional MW. Moreover, MMW can adjust a coefficient and prioritize relayed users in order to improve fairness, in terms of average delay, between the direct and relayed users. MMW can be employed both in centralized and decentralized network implementations, as well as the scenarios with shared or independent channels for the BS

and RS. In particular, in the case of independent channels for the BS and RS, by defining a suitably large threshold, a completely distributed resource allocation is possible most of the time without any signaling between the BS and RS about the QCSI of the relayed users. Numerical results confirm this as well as the fact that MMW is able to improve the delay fairness in the system and provides similar performance for direct and relayed users.

Chapter 7

Conclusions and Future Work

In this chapter, we conclude the presented works in this thesis and also suggest several topics for future work. Note that the conclusions provided in the following address the questions and objectives stated in Chapter 1. In particular, the conclusion for Chapter 2 addresses the question about the effect of buffer-aided relaying on the end-to-end delay. The service oriented concern for QoS provisioning in a multiuser system is addressed in the conclusion related to Chapter 3. Finally, the conclusions of Chapters 4, 5 and 6 address the implementation oriented concerns about the low-complexity and efficiency of the resource allocation algorithms.

7.1 Conclusions

- In Chapter 2, we have provided insights on the effect of buffer-aided relaying on the end-to-end delay, i.e., the delay that data packets experience since their arrival at the BS buffer until delivery to the destination. We have shown that using buffer at the relay helps in utilizing the BS channel more opportunistically and results in the fast transfer of the data packets from the BS buffer to the relay buffer. Then, the queued data in the relay buffer are served whenever the relay channel is in good condition. This way, the BS and relay channels are used more efficiently and the data packets are delivered to the destination in a less amount of time, compared with the conventional relaying. Also, we have analyzed the average packet delays

in the relay networks with Bernoulli packet arrivals and channel conditions, and we have shown mathematically that the average delay is less in the case of buffer-aided relaying. Furthermore, through intuitive generalizations, we have clarified that the throughput improvement in buffer-aided relaying in fact leads to the improvement in the average end-to-end delay performance as well. We have verified our analysis and discussions through extensive computer simulations. Numerical results confirm that even though buffer-aided relaying leads to queueing delays at the relay, it significantly reduces queueing delay at the BS and therefore, on the whole, it reduces average end-to-end packet delay. These results helps in deciding about the use of buffering relays in cellular networks as they dispel the concern on the delay performance of the system.

- In Chapter 3, we have presented a novel framework for formulating QoS-aware resource allocation problem in buffer-aided relay-enhanced OFDMA networks. We have shown that the IRAP formulation is itself a challenge when both the services with average throughput requirement and the services with packet delay thresholds are present in the network. To address that, we have proposed novel CQDA policies. Based on the approaches these policies have for problem formulation, we have called them SUMR and JUMR. SUMR defines the utility function based on only the delay-tolerant users and imposes minimum rate constraints whenever there are delay-sensitive packets in the queues of the BS and relays. On the other hand, JUMR defines utility function based on all the users and imposes minimum rate constraints only when getting close to the packet deadlines of delay-sensitive users. We have also proposed enhanced versions of these policies, as ESUMR and EJUMR, which use more information or computations in formulating or solving the IRAP. Through extensive computer simulations, we have evaluated the performances of the proposed

policies in extreme scenarios as well as general ones. The results show significant improvements in provisioning average throughput and packet delay guarantees, compared with the systems without relays, relay enhanced systems without buffers at the relays, as well as the systems with buffer-aided relays but QoS-unaware. Also, we have observed that when the system load increases, SUMR and ESUMR work in the favor of delay-sensitive users and keep their packet drop ratios equal to zero, at the cost of lowering the throughput of delay-tolerant users. On the other hand, JUMR and EJUMR are able to jointly serve the delay-sensitive and delay-tolerant users, and penalize both of them whenever the system load increases.

- In Chapter 4, we have presented a novel framework for distributed resource allocation in a buffer-aided relay-assisted OFDMA network. We have provided a new perspective, which considers the buffers at the relays as virtual users and models the whole network as small cells served by the BS and relays. Based on that and using the concept of time sharing, we have formulated the resource allocation problem as a convex optimization problem. Using dual decomposition, we have proposed an iterative algorithm called DDRA, which provides insights on reducing the computational burden on the BS and the CSI reporting overhead of the system. In DDRA, the BS and relays pass messages among themselves to solve their own problems using some global variables and the information about their queues and channel states of their links. The closed form equations for power and subchannel allocation in DDRA reveal that it adaptively allocates the system resources based on the queue sizes, channel conditions and required BER of the users. Numerical results confirm that DDRA is able to utilize the potential of buffer-aided relaying and results in significant improvement in terms of average throughput and queue stability.

- In Chapter 5, we have introduced important parameters that adapt the Lyapunov

drift-plus-penalty policy to cellular networks with buffering relays. In particular, the importance parameter for average power constraints amplifies the effect of virtual power queues of the BS and relays in the instantaneous problem. This way, it prevents continuous use of the peak powers at the BS and relays, and facilitates satisfying average power constraints in the presence of large actual data queue sizes. The other parameter is the extra weight that is given to the links of relayed users and increases their priority to enable more service for the queues of relayed users. As a consequence, it reduces the queue size of relayed users in the BS and helps in providing fair data admission for direct and relayed users. We have also proposed a low-complexity subchannel and power allocation strategy and based on that, we have designed distributed and centralized resource allocation schemes, respectively called EDDRA and EDCRA, which take into account several practical constraints such as limited buffer capacities and half-duplex relaying. EDDRA splits resource allocation tasks between the BS and relays and leads to lower computational burden on the BS and lower CSI overhead compared with EDCRA. In particular, in EDDRA, the BS decides about the type of time slot and the set of subchannels for the relays and itself. Then every one of them allocates its set of subchannels to its links in a distributed way and adjusts its total power to be used on the subchannels. Numerical results show that the proposed parameters lead to fair data admission for the users and help in satisfying the average power constraints. Also, they confirm that the EDDRA has close performance to EDCRA and outperforms an existing centralized algorithm in terms of system utility, overflow and throughput.

- In Chapter 6, we have proposed MMW policy which assigns the weight of a link based on the queue size in its starting point, unless in the rare events that the queue size in the ending point of a feeder link from the BS to relay, exceeds a predefined

large threshold, in which case its weight is defined according to conventional MW. MMW reduces the average delay for the relayed users. Moreover, by adjusting a coefficient, it is able to further prioritize the relayed users' links and improve the delay fairness between the direct and relayed users. MMW can be exploited both in centralized and decentralized network implementations as well as the scenarios with shared or independent channels for the BS and relays. In particular, when the BS and relays have dedicated channels, by defining a suitably large threshold, MMW leads to a completely distributed resource allocation without any signaling between the BS and relays about the local channel and queue states of the relayed users. Numerical results confirm that MMW causes lower overhead and improves delay fairness between direct and relayed users.

Note that the numerical results in this thesis are obtained from Monte Carlo simulations based on the well established models for random data arrivals and wireless channels, which are widely used in the literature. Therefore, even though these models may not match exactly the practical scenarios, the performance improvements of our proposed schemes over the baseline methods existing in the literature are expected to hold, as the same models have been used in simulating the performance of the proposed algorithms as well as the baseline methods.

7.2 Suggestions for Future Work

In the following, we consider several interesting possibilities for extension of the current work.

1. **Effect of Buffer-Aided Relaying on the Delay Variance:** Research on buffer-aided relaying is new, and more investigations are needed to analyze different per-

formance metrics in the presence of buffering relays. Several works in the literature [32, 35, 38, 56] as well as the investigations in this thesis have already shown that using buffers at relays leads to throughput improvements. Moreover, we have shown that buffer-aided relaying also reduces average end-to-end packet delays. However, the effect of buffer-aided relaying on the delay variance is an open research problem. It is needed to identify the different affecting factors that can lead to large or small variances of end-to-end packet delays.

2. **Energy Efficient Resource Allocation:** The growing demands for wireless access to the Internet entail a great amount of energy consumption in cellular networks, which inevitably leads to a bigger carbon footprint, and greatly contributes to environmental pollution. Therefore, considering Energy Efficiency (EE) in the system design is becoming an urgent trend, and recently remarkable efforts have been invested in this area [89–95]. In energy efficient design, the goal is to optimize the amount of data transmitted per unit energy, which requires tradeoffs for Spectral Efficiency (SE) and system capacity. Considering the fact that the buffering capability in relays improves the system capacity, it can result in better tradeoffs between EE and SE. This needs to be investigated more and the affecting parameters need to be identified. Moreover, since EE can lead to larger queue sizes in the system, it is necessary to study the instantaneous problem formulation such that the EE is optimized over time while keeping the queues stable.

3. **Mobility Awareness:** One of the challenges in cellular networks is maintaining wireless connectivity for the users that move in the network area, leaving the coverage of one serving node and entering a new one. This is usually addressed through link quality measurements and handover mechanisms. If the serving nodes are buffer-aided fixed relays, the data buffered in the previous relay needs to be retransmitted

from the BS to the new relay. In the scenarios that the users have high mobility, like driving in a road, this can affect the throughput performance of the system, as the volume of retransmissions is large. Therefore, suitable predictive resource allocation algorithms are needed to prevent the BS from transmitting to the old relay, early enough before the handover.

On the other hand, for the scenarios that a lot of users move together, like in a bus or train, mobile relays have been considered as a potential solution to reduce the overhead of handover signaling [96–100]. Using buffering capability in mobile relays can improve the reliability and quality of service in such scenarios. However, suitable predictive resource allocation methods are needed in this case, to feed the buffers of users in the mobile relay, early enough, to ensure a satisfactory service during the handover procedure.

4. **Multicell Scenarios:** One of the important research topics is to consider the challenges encountered in multicell systems. Noting that the next generation of cellular networks will be aggressive in frequency reusing, the interference resulted from adjacent cells can degrade the performance of the systems. There has been continuous work going on in this area [101–105]. However, these works either consider the networks without relays or the relay networks without buffering capability in relays. Considering the fact that the buffering capability allows to postpone data transmissions from relays, it can facilitate the interference management mechanisms. However, coordination between the adjacent cells and the involved entities is challenging as the queue states in relays should also be taken into account, in addition to the channel conditions. Therefore, resource allocation in buffer-aided relay-assisted multicell scenarios needs to be investigated and efficient algorithms need to be designed.

Bibliography

[1] D. Pareit, B. Lannoo, I. Moerman, and P. Demeester, "The history of WiMAX: A complete survey of the evolution in certification and standardization for IEEE 802.16 and WiMAX," *IEEE Commun. Surveys and Tutorials*, vol. 14, no. 4, pp. 1183–1211, Oct. 2012.

[2] "IEEE standard for air interface for broadband wireless access systems," IEEE Std. 802.16-2012 (pp. 12544) (2012).

[3] S. Parkvall, A. Furuskar, and E. Dahlman, "Evolution of LTE toward IMT-advanced," *IEEE Commun. Magazine*, vol. 49, no. 2, pp. 84–91, Feb. 2011.

[4] "Evolved universal terrestrial radio access (E-UTRA) and evolved universal terrestrial radio access networks(E-UTRAN; overall description: Stage 2," 3GPP, TS 36.300, Rel. 12, Mar. 2015.

[5] J. Son, "New feature for IMT-advanced: Relay," *Samsung Electronics Training Workshop on 4G Mobile (IMT Advanced) System and Applications*, Nov. 2009. [Online]. Available: https://www.itu.int/ITU-D/asp/CMS/Events/2009/CoE/4Gmobile/Session6_SON.pdf

[6] C. Hoymann, W. Chen, J. Montojo, A. Golitschek, C. Koutsimanis, and X. Shen, "Relaying operation in 3GPP LTE: challenges and solutions," *IEEE Commun. Magazine*, vol. 50, no. 2, pp. 156–162, Feb. 2012.

[7] M. Salem, A. Adinoyi, M. Rahman, H. Yanikomeroglu, D. Falconer, Y. Kim, E. Kim, and Y. Cheong, "An overview of radio resource management in relay-enhanced OFDMA-based networks," *IEEE Commun. Surveys and Tutorials*, vol. 12, no. 3, pp. 422–438, Third quarter 2010.

[8] M. Salem, A. Adinoyi, H. Yanikomeroglu, and D. Falconer, "Opportunities and challenges in OFDMA-based cellular relay networks: A radio resource management perspective," *IEEE Trans. Veh. Technol.*, vol. 59, no. 5, pp. 2496–2510, June 2010.

[9] K. Sundaresan and S. Rangarajan, "Adaptive resource scheduling in wireless OFDMA relay networks," in *Proc. IEEE Conf. on Computer Commun.*, Mar. 2012, pp. 1080–1088.

[10] D. Ng and R. Schober, "Cross-layer scheduling for OFDMA amplify-and-forward relay networks," *IEEE Trans. Veh. Technol.*, vol. 59, no. 3, pp. 1443–1458, Mar. 2010.

[11] D. Ng, E. Lo, and R. Schober, "Dynamic resource allocation in MIMO-OFDMA systems with full-duplex and hybrid relaying," *IEEE Trans. Commun.*, vol. 60, no. 5, pp. 1291–1304, May 2012.

[12] Y. Pan, A. Nix, and M. Beach, "Distributed resource allocation for OFDMA-based relay networks," *IEEE Trans. Veh. Technol.*, vol. 60, no. 3, pp. 919–931, Mar. 2011.

[13] S. Zhang and X. Xia, "A high-efficiency semi-distributed resource allocation in OFDMA-based wireless relay networks," in *Proc. IEEE Wireless Commun. and Networking Conf.*, Apr. 2013, pp. 3277–3281.

[14] Z. Chang, T. Ristaniemi, and Z. Niu, "Radio resource allocation for collaborative OFDMA relay networks with imperfect channel state information," *IEEE Trans. Wireless Commun.*, vol. 13, no. 5, pp. 2824–2835, May 2014.

[15] O. Oyman, "Opportunistic scheduling and spectrum reuse in relay-based cellular networks," *IEEE Trans. Wireless Commun.*, vol. 9, no. 3, pp. 1074–1085, Mar. 2010.

[16] J. Liang, H. Yin, H. Chen, Z. Li, and S. Liu, "A novel dynamic full frequency reuse scheme in OFDMA cellular relay networks," in *Proc. IEEE Veh. Technol. Conf.*, Sep. 2011, pp. 1–5.

[17] H. Mei, J. Bigham, P. Jiang, and E. Bodanese, "Distributed dynamic frequency allocation in fractional frequency reused relay based cellular networks," *IEEE Trans. Commun.*, vol. 61, no. 4, pp. 1327–1336, Apr. 2013.

[18] W. Jeon, J. Han, and D. Jeong, "Distributed resource allocation for multi-cell relay-aided OFDMA systems," *IEEE Trans. Mobile Computing*, vol. 13, pp. 2003–2015, Sep. 2014.

[19] L. Jiang, J. Pang, G. Shen, and D. Wang, "A game theoretic channel allocation scheme for multi-user OFDMA relay system," in *Proc. IEEE Wireless Commun. and Networking Conf.*, Mar. 2011, pp. 298–303.

[20] L. Liang and G. Feng, "A game-theoretic framework for interference coordination in OFDMA relay networks," *IEEE Trans. Veh. Technol.*, vol. 61, no. 1, pp. 321–332, Jan. 2012.

[21] I. Chaabane, S. Hamouda, S. Tabbane, and J. Vicario, "A new PRB sharing scheme in dual-hop LTE-advanced system using game theory," in *Proc. IEEE Personal, Indoor and Mobile Radio Commun. Sympos.*, Sep. 2012, pp. 375–379.

[22] Y. Farazmand and A. Alfa, "A game theoretic power allocation and relay load balancing in OFDMA-based DF cellular relay networks," in *Proc. IEEE Veh. Technol. Conf.*, Sep. 2013, pp. 1–6.

[23] D. Zhang, Y. Wang, and J. Lu, "QoS-aware relay selection and subcarrier allocation in cooperative OFDMA systems," *IEEE Commun. Lett.*, vol. 14, no. 4, pp. 294–296, Apr. 2010.

[24] D. Zhang, X. Tao, J. Lu, and M. Wang, "Dynamic resource allocation for real-time services in cooperative OFDMA systems," *IEEE Commun. Lett.*, vol. 15, no. 5, pp. 497–499, May 2011.

[25] X. Zhang, X. Tao, Y. Li, and J. Lu, "QoS provisioning scheduling with joint optimization of base station and relay power allocation in cooperative OFDMA systems," in *Proc. IEEE Intern. Commun. Conf.*, June 2013, pp. 5453–5457.

[26] A. Marques, C. Figuera, C. Rey-Moreno, and J. Simo-Reigadas, "Asymptotically optimal cross-layer schemes for relay networks with short-term and long-term constraints," *IEEE Trans. Wireless Commun.*, vol. 12, pp. 333–345, Jan. 2013.

[27] O. Elgendy, M. Ismail, and K. Elsayed, "Max-min fair resource allocation for LTE-advanced relay-enhanced cells," in *Proc. IEEE Wireless Commun. and Networking Conf.*, Apr. 2014, pp. 1432–1437.

[28] Y. Farazmand and A. Alfa, "Power allocation framework for OFDMA-based decode-and-forward cellular relay networks," *IEEE J. Commun. and Networks*, vol. 16, pp. 559–567, Oct. 2014.

[29] Y. Zhao, X. Fang, R. Huang, and Y. Fang, "Joint interference coordination and load balancing for OFDMA multihop cellular networks," *IEEE Trans. Mobile Computing*, vol. 13, pp. 89–101, Jan. 2014.

[30] D. Incebacak, H. Yanikomeroglu, and B. Tavli, "Trade-offs in sum-rate maximization and fairness in relay-enhanced OFDMA-based cellular networks," in *Proc. IEEE Global Telecommun. Conf.*, Dec. 2014, pp. 4770–4775.

[31] C. Yangyang, P. Martins, L. Decreusefond, F. Yan, and X. Lagrange, "Stochastic analysis of a cellular network with mobile relays," in *Proc. IEEE Global Telecommun. Conf.*, Dec. 2014, pp. 4758–4763.

[32] N. Mehta, V. Sharma, and G. Bansal, "Performance analysis of a cooperative system with rateless codes and buffered relays," *IEEE Trans. Wireless Commun.*, vol. 10, no. 4, pp. 1069–1081, April 2011.

[33] N. Zlatanov, R. Schober, and P. Popovski, "Throughput and diversity gain of buffer-aided relaying," in *Proc. IEEE Global Telecommun. Conf.*, Dec. 2011, pp. 1–6.

[34] C. Dong, L. Yang, and L. Hanzo, "Performance analysis of multihop-diversity-aided multihop links," *IEEE Trans. Veh. Technol.*, vol. 61, no. 6, pp. 2504–2516, July 2012.

[35] N. Zlatanov and R. Schober, "Buffer-aided relaying with adaptive link selection-fixed and mixed rate transmission," *IEEE Trans. Inform. Theory*, vol. 59, no. 5, pp. 2816–2840, May 2013.

[36] N. Zlatanov, R. Schober, and P. Popovski, "Buffer-aided relaying with adaptive link selection," *IEEE J. Select. Areas Commun.*, vol. 31, no. 8, pp. 1530–1542, Aug. 2013.

[37] V. Jamali, N. Zlatanov, A. Ikhlef, and R. Schober, "Achievable rate region of the bidirectional buffer-aided relay channel with block fading," *IEEE Trans. Inform. Theory*, vol. 60, no. 11, pp. 7090–7111, Nov. 2014.

[38] N. Zlatanov, A. Ikhlef, T. Islam, and R. Schober, "Buffer-aided cooperative communications: opportunities and challenges," *IEEE Commun. Magazine*, vol. 52, no. 4, pp. 146–153, Apr. 2014.

[39] I. Krikidis, T. Charalambous, and J. Thompson, "Buffer-aided relay selection for cooperative diversity systems without delay constraints," *IEEE Trans. Wireless Commun.*, vol. 11, no. 5, pp. 1957–1967, May 2012.

[40] T. Islam, A. Ikhlef, R. Schober, and V. Bhargava, "Multisource buffer-aided relay networks: Adaptive rate transmission," in *Proc. IEEE Global Telecommun. Conf.*, Dec. 2013, pp. 3577–3582.

[41] H. Liu, P. Popovski, E. de Carvalho, and Y. Zhao, "Sum-rate optimization in a two-way relay network with buffering," *IEEE Commun. Lett.*, vol. 17, no. 1, pp. 95–98, Jan. 2013.

[42] I. Ahmed, A. Ikhlef, R. Schober, and R. Mallik, "Power allocation for conventional and buffer-aided link adaptive relaying systems with energy harvesting nodes," *IEEE Trans. Wireless Commun.*, vol. 13, no. 3, pp. 1182–1195, Mar. 2014.

[43] J. Huang and A. Swindlehurst, "Wireless physical layer security enhancement with buffer-aided relaying," in *Asilomar Conf. on Signals, Systems and Computers*, Nov. 2013, pp. 1560–1564.

[44] ——, "Buffer-aided relaying for two-hop secure communication," *IEEE Trans. Wireless Commun.*, vol. 14, no. 1, pp. 152–164, Jan. 2015.

[45] M. Darabi, V. Jamali, B. Maham, and R. Schober, "Adaptive link selection for cognitive buffer-aided relay networks," *IEEE Commun. Lett.*, vol. 19, no. 4, pp. 693–696, Apr. 2015.

[46] D. Park, "A throughput-optimal scheduling policy for wireless relay networks," in *Proc. IEEE Wireless Commun. and Networking Conf.*, Apr. 2010, pp. 1–5.

[47] M. Salem, A. Adinoyi, M. Rahman, H. Yanikomeroglu, D. Falconer, and Y. Kim, "Fairness-aware radio resource management in downlink OFDMA cellular relay networks," *IEEE Trans. Wireless Commun.*, vol. 9, no. 5, pp. 1628–1639, May 2010.

[48] M. Salem, A. Adinoyi, H. Yanikomeroglu, and D. Falconer, "Fair resource allocation toward ubiquitous coverage in OFDMA-based cellular relay networks with asymmetric traffic," *IEEE Trans. Veh. Technol.*, vol. 60, no. 5, pp. 2280–2292, June 2011.

[49] I. Bastuerk, B. Oezbek, and D. Ruyet, "Queue-aware resource allocation for OFDMA-based mobile relay enhanced networks," in *Proc. Intern. Sympos. on Wireless Commun. Systems*, Aug. 2013, pp. 1–5.

[50] R. Wang and V. Lau, "Delay-aware two-hop cooperative relay communications via approximate MDP and stochastic learning," *IEEE Trans. Inform. Theory*, vol. 59, no. 11, pp. 7645–7670, Nov 2013.

[51] H. Ju, B. Liang, J. Li, and X. Yang, "Dynamic joint resource optimization for LTE-Advanced relay networks," *IEEE Trans. Wireless Commun.*, vol. 12, no. 11, pp. 5668–5678, Nov. 2013.

[52] L. Wang, Q. Du, P. Ren, L. Sun, and Y. Wang, "Buffering-aided resource allocation for type I relay in LTE-advanced cellular networks," in *Proc. IEEE Global Telecommun. Conf.*, Dec. 2014, pp. 4484–4489.

[53] B. Zhou, Y. Cui, and M. Tao, "Stochastic throughput optimization for two-hop systems with finite relay buffers," in *Proc. IEEE Global Telecommun. Conf.*, Dec. 2014, pp. 1728–1733.

[54] L. Georgiadis, M. Neely, and L. Tassiulas, "Resource allocation and cross-layer control in wireless networks," *Foundation and Trends in Networking*, vol. 1, no. 1, pp. 1–144, Apr. 2006.

[55] M. Neely, *Stochastic Network Optimization with Application to Communication and Queueing Systems*. Morgan & Claypool, 2010.

[56] B. Xia, Y. Fan, J. Thompson, and H. Poor, "Buffering in a three-node relay network," *IEEE Trans. Wireless Commun.*, vol. 7, no. 11, pp. 4492–4496, Nov. 2008.

[57] F. Gebali, *Analysis of Computer and Communication Networks*. Springer, 2008.

[58] I. Adan and J. Resing, *Queueing Systems*. Eindhoven University: Online-Available: http://www.win.tue.nl/ iadan/queueing.pdf, 2015.

[59] M. Andrews, K. Kumaran, K. Ramanan, A. Stolyar, P. Whiting, and R. Vijayakumar, "Providing quality of service over a shared wireless link," *IEEE Commun. Magazine*, vol. 39, pp. 150–153, 2001.

[60] M. Neely, E. Modiano, and C. Rohrs, "Dynamic power allocation and routing for time varying wireless networks," *IEEE J. Select. Areas Commun.*, vol. 23, no. 1, pp. 89–103, Jan. 2005.

[61] "Spatial channel model for multiple input multiple output (MIMO) simulations," 3GPP TR 25.996 V7.0.0 (2007-06).

[62] M. Jeruchim, P. Balaban, and K. Shanmugan, *Simulation of Communication Systems: Modeling, Methodology and Techniques*, 2nd ed. Kluwer Academic, 2000.

[63] X. Qiu and K. Chawla, "On the performance of adaptive modulation in cellular systems," *IEEE Trans. Commun.*, vol. 47, no. 6, pp. 884–895, June 1999.

[64] Y. Kim, K. Son, and S. Global, "QoS scheduling for heterogeneous traffic in OFDMA-based wireless systems," in *Proc. IEEE Global Telecommun. Conf.*, Nov. 2009, pp. 1–6.

[65] C. Wanshi, M. Neely, and U. Mitra, "Energy efficient scheduling with individual packet delay constraints: Offline and online results," in *Proc. IEEE Conf. on Computer Commun.*, May 2007, pp. 1136–1144.

[66] J. Gan, Z. Guo, F. Rui, L. Weihong, W. Hai, K. Sandlund, L. Jianjun, S. Xiaodong, and L. Guangyi, "LTE in-band relay prototype and field measurement," in *Proc. IEEE Veh. Technol. Conf.*, May 2012, pp. 1–5.

[67] L. Wang, Y. Ji, and F. Liu, "A semi-distributed resource allocation scheme for OFDMA relay-enhanced downlink systems," in *Proc. IEEE Global Telecommun. Conf.*, Dec. 2008, pp. 1–6.

[68] J. Huang, V. Subramanian, R. Agrawal, and R. Berry, "Downlink scheduling and resource allocation for OFDM systems," *IEEE Trans. Wireless Commun.*, vol. 8, no. 1, pp. 288–296, Jan. 2009.

[69] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge University Press, 2004.

[70] D. Bertsekas, *Nonlinear Programming*, 2nd ed. Athena Scientific, 1999.

[71] L. Huang, M. Rong, L. Wang, Y. Xue, and E. Schulz, "Resouce scheduling for OFDMA/TDD based relay enhanced cellular networks," in *Proc. IEEE Wireless Commun. and Networking Conf.*, Mar. 2007, pp. 1544–1548.

[72] B. Bangerter, S. Talwar, R. Arefi, and K. Stewart, "Networks and devices for the 5G era," *IEEE Commun. Magazine*, vol. 52, pp. 90–96, Feb. 2014.

[73] L. Tassiulas and A. Ephremides, "Stability properties of constrained queueing systems and scheduling policies for maximum throughput in multihop radio networks," *IEEE Trans. Automatic Control*, vol. 37, no. 12, pp. 1936–1948, 1992.

[74] ——, "Dynamic server allocation to parallel queues with randomly varying connectivity," *IEEE Trans. Inform. Theory*, vol. 39, pp. 466–478, 1993.

[75] "Evolved universal terrestrial radio access (E-UTRA); further advancements for E-UTRA physical layer aspects (release 9)," TR 36.814 V9.0.0 (2010-03).

[76] P. Bender, P. Black, M. Grob, R. Padovani, N. Sindhushyana, and S. Viterbi, "CDMA/HDR: a bandwidth efficient high speed wireless data service for nomadic users," *IEEE Commun. Magazine*, vol. 38, pp. 70–77, July 2000.

[77] M. Andrews and L. Zhang, "Scheduling algorithms for multi-carrier wireless data systems," in *ACM Intern. Conf. on Mobile Computing and Networking*, Sep. 2007.

[78] G. Song and Y. Li, "Cross-layer optimization for OFDM wireless networks-part ii: Algorithm development," *IEEE Trans. Wireless Commun.*, vol. 4, no. 2, pp. 625–634, Mar. 2005.

[79] ——, "Utility-based resource allocation and scheduling in OFDM-based wireless broadband networks," *IEEE Commun. Magazine*, vol. 43, no. 12, pp. 127–134, Dec. 2005.

[80] Y. Ma, "Rate-maximization scheduling for downlink OFDMA with long term rate proportional fairness," in *Proc. IEEE Intern. Conf. on Commun.*, May 2008, pp. 3480–3484.

[81] ——, "Rate maximization for downlink OFDMA with proportional fairness," *IEEE Trans. Veh. Technol.*, vol. 57, no. 5, pp. 3267–3274, Sep. 2008.

[82] C. Mohanram and S. Bhashyam, "Joint subcarrier and power allocation in channel-aware queue-aware scheduling for multiuser OFDM," *IEEE Trans. Wireless Commun.*, vol. 6, pp. 3208–3213, Sep. 2007.

[83] G. Song, Y. Li, and L. Cimini, "Joint channel-and queue-aware scheduling for multiuser diversity in wireless OFDMA networks," *IEEE Trans. Commun.*, vol. 57, no. 7, pp. 2109–2121, July 2009.

[84] H. Zhu and J. Wang, "Chunk-based resource allocation in OFDMA systems - part i: Chunk allocation," *IEEE Trans. Commun.*, vol. 57, no. 9, pp. 2734–2744, Sep. 2009.

[85] ——, "Chunk-based resource allocation in OFDMA systems-part ii: Joint chunk, power and bit allocation," *IEEE Trans. Commun.*, vol. 60, no. 2, pp. 499–509, Feb. 2012.

[86] E. Larsson, "Optimal OFDMA downlink scheduling under a control signaling cost constraint," *IEEE Trans. Commun.*, vol. 58, pp. 2776 – 2781, Sep. 2010.

[87] J. Little, "A proof for the queuing formula: L=λW," *Operations Research*, vol. 9, no. 3, pp. 383–387, June 1961.

[88] R. Jain, D. Chiu, and W. Hawe, "A quantitative measure of fairness and discrimination for resource allocation in shared computer systems," *DEC Research Report TR-301*, Sep. 1984.

[89] C. Xiong, G. Li, S. Zhang, Y. Chen, and S. Xu, "Energy-efficient resource allocation in OFDMA networks," in *Proc. IEEE Global Telecommun. Conf.*, Dec. 2011, pp. 1–5.

[90] C. Ho and C. Huang, "Energy-efficient 2-D resource allocation with fairness constraints for OFDMA networks," in *Proc. IEEE Veh. Technol. Conf.*, Sep. 2011, pp. 1–5.

[91] S. Huang, H. Chen, J. Cai, and F. Zhao, "Energy efficiency and spectral-efficiency tradeoff in amplify-and-forward relay networks," *IEEE Trans. Veh. Technol.*, vol. 62, no. 9, pp. 4366–4378, Nov. 2013.

[92] I. Ku, C. Wang, and J. Thompson, "Spectral-energy efficiency tradeoff in relay-aided cellular networks," vol. 12, no. 10, pp. 4970–4982, Oct. 2013.

[93] F. Parzysz, M. Vu, and F. Gagnon, "Trade-offs on energy-efficient relay deployment in cellular networks," in *Proc. IEEE Veh. Technol. Conf.*, Sep. 2014, pp. 1–6.

[94] Y. Chen, X. Fang, and B. Huang, "Energy-efficient relay selection and resource allocation in nonregenerative relay OFDMA systems," *IEEE Trans. Veh. Technol.*, vol. 63, no. 8, pp. 3689–3699, Oct. 2014.

[95] L. Venturino, A. Zappone, C. Risi, and S. Buzzi, "Energy-efficient scheduling and power allocation in downlink OFDMA networks with base station coordination," *IEEE Trans. Wireless Commun.*, vol. 14, no. 1, pp. 1–14, Jan. 2015.

[96] R. Balakrishnan, X. Yang, M. Venkatachalam, and I. Akyildiz, "Mobile relay and group mobility for 4G WiMAX networks," in *Proc. IEEE Wireless Commun. and Networking Conf.*, Mar. 2011, pp. 1224–1229.

[97] L. Chen, Y. Huang, F. Xie, Y. Gao, L. Chu, H. He, Y. Li, F. Liang, and Y. Yuan, "Mobile relay in LTE-advanced systems," *IEEE Commun. Magazine*, vol. 51, no. 11, pp. 144–151, Nov. 2013.

[98] H. Zhao, R. Huang, J. Zhang, and Y. Fang, "Handoff for wireless networks with mobile relay stations," in *Proc. IEEE Wireless Commun. and Networking Conf.*, Mar. 2011, pp. 826–831.

[99] H. Chang, M. Ku, K. Singh, and J. Lin, "Low-complexity amplify-and-forward mobile relay networks without source-to-relay CSI," in *Proc. IEEE Veh. Technol. Conf.*, Sep. 2013, pp. 1–5.

[100] Z. Liao, X. Zhang, and C. Feng, "Mobile relay deployment based on Markov chains in WiMAX networks," in *Proc. IEEE Global Telecommun. Conf.*, Dec. 2014, pp. 4508–4513.

[101] J. Eun, H. Shin, and J. Lee, "Inter-cell interference coordination for a downlink OFDMA relay network with multicells," in *Proc. IEEE Veh. Technol. Conf.*, May 2012, pp. 1–5.

[102] A. Argyriou, "Interference decoding in cellular wireless relay networks with space-time coding," in *Proc. IEEE Wireless Commun. and Networking Conf.*, Apr. 2014, pp. 1160–1165.

[103] A. Omri and M. Hasna, "Performance analysis of OFDMA based wireless cooperative networks with interference management," in *Proc. IEEE Veh. Technol. Conf.*, June 2013, pp. 1–6.

[104] Y. Zhao, X. Fang, R. Huang, and Y. Fang, "Joint interference coordination and load balancing for OFDMA multihop cellular networks," *IEEE Trans. Mobile Computing*, vol. 13, no. 1, pp. 89–101, Jan. 2014.

[105] M. Fallgren, "An optimization approach to joint cell, channel and power allocation in multicell relay networks," *IEEE Trans. Wireless Commun.*, vol. 11, no. 8, pp. 2868–2875, Aug. 2012.

[106] "Evolved universal terrestrial radio access (E-UTRA); physical channels and modulation," 3GPP, TS 36.211, Rel. 12, Mar. 2015.

Appendix A

Assumptions for Channel Models

In cellular networks, data transmissions between different nodes go through wireless channels. Wireless channel affects a transmitted signal in different ways and therefore, when a signal arrives at the receiver, its power is different from the power at the transmitter. This is firstly due to the path loss, which is the reduction in the signal's power as it travels through the air. Second factor affecting the signal's power is the propagation of signal through multiple paths before reaching the receiver. This leads to the arrival of several copies of the signal at the receiver, with different attenuation and delays, which might add either constructively or destructively at the receiving antenna. This affect is called multipath fading [62, Chapter 9].

To simulate the behavior of wireless channels and their effect on the transmission rates of the links in the performance evaluations of this thesis, we have assumed the following:

- **The network environment is urban Macrocell [61].**

  The reason for this assumption is that using relays in cellular networks makes it possible to have large coverage area for a single cell served through the BS and relays. Considering this assumption, we have used the COST231 Hata urban propagation model as in [61] and computed the path loss of a link based on the following equation:

$$\begin{aligned} PL &= (44.9 - 6.55\log_{10}(h_{tx}))\log_{10}(\frac{d}{1000}) + 45.5 \\ &+ (35.46 - 1.1h_{rx})\log_{10}(f_c) - 13.82\log_{10}(h_{tx}) + 0.7h_{rx} + 3, \end{aligned} \quad \text{(A.1)}$$

where PL is the path loss in dB, h_{tx} is the transmitter antenna height in meters, h_{rx} the receiver antenna height in meters, f_c the carrier frequency in MHz, and d is the distance between the transmitter and receiver in meters.

- **The BS and relays are located in a way that there is Line of Sight (LOS) between them.**

  This assumption usually holds when the BS and relays' antennas are installed in high locations, as in our simulations. In this case, there is a dominant component among the copies of a signal that travel multiple paths from the BS to a relay. Considering this assumption, for the links between the BS and relays, we have used Rician channel model [61, Chapter 9].

- **The links between the BS and users as well as the links between the relays and users are in Non-Line of Sight (NLOS) condition.**

  This assumption usually holds as there are many blocks between the antennas of the BS/relays and the user equipment's antenna. Therefore, there is not a single dominant component contributing to the power of the signal arrived at the user equipment antenna. Considering this assumption, for the links between the BS/relays and users, we have used Rayleigh channel model [61, Chapter 9].

- **The channel conditions of all the links remain constant over a time slot but vary from one slot to another.**

  In our system models, we consider the scenarios in which the users are either fixed or have very limited mobility and the network environment has little changes over time. In this case, channel variations happen over time intervals larger than the symbol length of the transmitted signal, which is referred to as slow fading [61, Chapter 9]. Since the time slots considered in the simulations have large durations (1 millisecond)

compared to the symbol length of OFDM systems (which are usually in the order of less than 100 microsecond [2, 106]), this assumption matches the slow fading present in the scenarios considered.

- **The channel conditions of all the links are constant over a subchannel but are different from one subchannel to another.**

  We have assumed that the network environments do not have diverse distribution of the reflecting blocks, and multiple copies of a signal arrive at the receiver with small delay spread [61, Chapter 9]; therefore, the variations over frequency happen over larger bandwidths than the OFDM subcarrier frequency. In our simulations, we have considered subchannels composed of one or more OFDM subcarriers, which matches this behavior.

We acknowledge the fact that, in reality, the channel behaviors in cellular networks might be different from the models considered in our simulations. However, considering the fact that the same models have been used in simulating the behavior of the baseline algorithms existing in the literature, the relative performance improvements of our proposed algorithms are expected to hold.

Appendix B

Proof of Theorem 2.1

In order to prove that the buffer aided relaying system incurs equal or lower delay compared the conventional one, it is required to prove $E(D_{nb}) - E(D_b) \geq 0$. To show this, note that

$$E(D_{nb}) - E(D_b) = 1 + \frac{1-a}{s_1 s_2 - a} - \frac{1-a}{s_1 - a} - \frac{1-a}{s_2 - a} \tag{B.1}$$

By adding and subtracting the term $\frac{1-a}{s_1-a}\frac{1-a}{s_2-a}$ and rearranging the equations, we have

$$\begin{aligned} E(D_{nb}) - E(D_b) &= \frac{1-a}{s_1-a}\frac{1-a}{s_2-a} - \frac{1-a}{s_1-a} - \frac{1-a}{s_2-a} + 1 + \frac{1-a}{s_1 s_2 - a} - \frac{1-a}{s_1-a}\frac{1-a}{s_2-a} \\ &= \left(\frac{1-a}{s_1-a} - 1\right)\left(\frac{1-a}{s_2-a} - 1\right) + \frac{1-a}{s_1 s_2 - a} - \frac{1-a}{s_1-a}\frac{1-a}{s_2-a}. \end{aligned} \tag{B.2}$$

Since $0 < a < s_i$, $i = 0, 1$, and $s_i \leq 1$, we have $\frac{1-a}{s_i-a} \geq 1$, $i = 0, 1$. Therefore, the first term in the right hand side of Equation (B.2) is non-negative. Hence, it suffices to show

$$\frac{1-a}{s_1 s_2 - a} \geq \frac{1-a}{s_1-a}\frac{1-a}{s_2-a}. \tag{B.3}$$

By canceling $1 - a$ and cross-multiplying in Equation (B.3), we obtain

$$(s_1 - a)(s_2 - a) \geq (1 - a)(s_1 s_2 - a). \tag{B.4}$$

After multiplying both sides out and canceling the common terms of Equation (B.4),

we have:

$$a\,(1 - s_1)\,(1 - s_2) \geq 0, \tag{B.5}$$

which is always true since $s_1 \leq 1$ and $s_2 \leq 1$.

Appendix C

Proof of Theorem 6.1

Let define $\rho_{k_l} = \alpha_{k_l}\beta_{k_l}$, $\rho^{max} = \sup_l \rho_{k_l}$, $a^{max} = \sup_{l,t} a_{k_l}^{b_l}(t)$ and $c^{max} = \sup_{l,t} c_l(t)$. For a given number of M time slots (the use of which will be clarified later), consider a time slot t_0 at which the system queue sizes have grown large such that $Q_{k_l}^{b_l}(\tau) \geq c^{max}$, $l \in \mathcal{L}, t_0 \leq \tau \leq t_0 + M - 1$. Then we have $\min(Q_{k_l}^{b_l}(\tau), c_l(\tau)) = c_l(\tau)$, $l \in \mathcal{L}, t_0 \leq \tau \leq t_0 + M - 1$. Considering the definition (6.2), in these time slots the MMW objective (6.7) is equivalent to

$$\text{Maximize} \sum_{l\in\mathcal{L}} \rho_{k_l} w_l^{MMW}(\tau) r_l(\tau) \tag{C.1}$$

In this case, if all of the RS queues are larger than the threshold, MMW weights are similar to those of MDB and therefore, it will stabilize the system [60]. On the other hand, if any RS queue is less than $\hat{Q}$ (in which case the weights of the corresponding links from the BS would be considered proportional to the queue sizes in the BS, and not the difference of the queue sizes in the BS and RS), stability of the whole system can be proved by using M-slot Lyapunov drift, similar to the stability proof of MDB in [60]. The proof exploits Lyapunov drift theory and has the following structure: Based on the queueing system state $\underline{Q}(t_0)$, a Lyapunov function, $L(\underline{Q}(t_0))$, is defined as a measure of the queue congestion in the system, and an M-slot conditional Lyapunov drift, $\Delta_M(\underline{Q}(t_0))$, is defined as $\Delta_M(\underline{Q}(t_0)) = E\{L(\underline{Q}(t_0 + M)) - L(\underline{Q}(t_0))|\underline{Q}(t_0)\}$. It is shown in [60] that if the sufficient condition for stability holds, there exists a finite M and a stationary randomized algorithm, referred to as STAT, such that the M-slot drift for it satisfies

$\Delta_M^{STAT}(\underline{Q}(t_0)) \leq B - \epsilon g(\underline{Q}(t_0))$, where B is a finite value and $g(\underline{Q})$ is a non-negative increasing function of queue sizes in the system. This means that using STAT, the system's overall queue congestion tends to decrease if it becomes sufficiently large. The stability of MMW can be proved by comparing it with STAT and showing that the MMW provides a similar bound on the drift as well. This is shown in detail in the following.

Based on [60], we define the Lyapunov function as:

$$L(\underline{Q}(\tau)) = \sum_{l \in \mathcal{L}} \rho_{k_l} (Q_{k_l}^{b_l}(\tau))^2, \tag{C.2}$$

Based on (6.3), the following holds:

$$Q_{k_l}^{b_l}(t_0 + M) \leq \max[Q_{k_l}^{b_l}(t_0) - \sum_{\tau=t_0}^{t_0+M-1} r_{k_l}^{b_l}(\tau), 0] + \sum_{\tau=t_0}^{t_0+M-1} a_{k_l}^{b_l}(\tau) \tag{C.3}$$

The inequality is due to the fact that some of the arrived data may depart in the interval from t_0 to $t_0 + M - 1$. Let define the average transmission rate and average arrival rate over M-slot interval, as:

$$\tilde{r}_{k_l}^{b_l}(t_0) = \frac{1}{M} \sum_{\tau=t_0}^{t_0+M-1} r_{k_l}^{b_l}(\tau) \tag{C.4}$$
$$\tilde{a}_{k_l}^{b_l}(t_0) = \frac{1}{M} \sum_{\tau=t_0}^{t_0+M-1} a_{k_l}^{b_l}(\tau)$$

Considering (C.3) and (C.4), we have:

$$Q_{k_l}^{b_l}(t_0 + M) \leq \max[Q_{k_l}^{b_l}(t_0) - M\tilde{r}_{k_l}^{b_l}(t_0), 0] + M\tilde{a}_{k_l}^{b_l}(t_0) \tag{C.5}$$

By squaring both sides and noting that $(\max\{x,0\})^2 \leq x^2$, we get:

$$(Q_{k_l}^{b_l}(t_0+M))^2-(Q_{k_l}^{b_l}(t_0))^2 \leq M^2[(\tilde{r}_{k_l}^{b_l}(t_0))^2+(\tilde{a}_{k_l}^{b_l}(t_0))^2]-2MQ_{k_l}^{b_l}(t_0)[\tilde{r}_{k_l}^{b_l}(t_0)-\tilde{a}_{k_l}^{b_l}(t_0)] \quad \text{(C.6)}$$

Note that the data arrivals at the RS queues are always less than or equal to the transmission rates from the corresponding queues in the BS, i.e. $a_k^R(t) \leq r_k^B(t)$. Considering this, multiplying both sides by ρ_{k_l}, summing over all the links and taking conditional expectation, results in the following inequality for the M-slot Lyapunov drift by any algorithm X:

$$\Delta_M^X = \sum_{l\in\mathcal{L}} \rho_{k_l} E\{(Q_{k_l}^{b_l}(t_0+M))^2-(Q_{k_l}^{b_l}(t_0))^2|\underline{Q}(t_0)\} \leq B-2M[\Phi^X(\underline{Q}(t_0))-\Theta(\underline{Q}(t_0))] \quad \text{(C.7)}$$

where $\Phi^X(\underline{Q}(t_0))$ and $\Theta(\underline{Q}(t_0))$ are defined as:

$$\begin{aligned}
\Theta(\underline{Q}(t_0)) &= E\{\sum_{l\in\mathcal{L}^B} \rho_{k_l} Q_{k_l}^B(t_0)\tilde{a}_{k_l}^B(t_0)|\underline{Q}(t_0)\} = E\{\sum_{k\in\mathcal{K}} \rho_k Q_k^B(t_0)\tilde{a}_k^B(t_0)|\underline{Q}(t_0)\} \\
\Phi^X(\underline{Q}(t_0)) &= E\{\sum_{l\in\mathcal{L}^B} \rho_{k_l} Q_{k_l}^B(t_0)\tilde{r}_{k_l}^{B(X)}(t_0) + \sum_{l\in\mathcal{L}^R} \rho_{k_l} Q_{k_l}^R(t_0)[\tilde{r}_{k_l}^{R(X)}(t_0)-\tilde{r}_{k_l}^{B(X)}(t_0)]|\underline{Q}(t_0)\}
\end{aligned} \quad \text{(C.8)}$$

and B is a constant such that $M^2\sum_{l\in\mathcal{L}} \rho_{k_l} E\{(\tilde{r}_{k_l}^{b_l}(t_0))^2+(\tilde{a}_{k_l}^{b_l}(t_0))^2|\underline{Q}(t_0)\} \leq B$, which exists due to the fact that channel capacities and the packet arrivals have finite mean and variance.

Based on [60], for an $\epsilon > 0$, there exists a finite M and a stationary randomized scheduling algorithm, STAT, which satisfies the following:

$$\begin{aligned}
E\{\tilde{r}_k^{B(STAT)}(t_0)|\underline{Q}(t_0)\} - E\{\tilde{a}_k^B(t_0)|\underline{Q}(t_0)\} \geq \frac{\epsilon}{2}, \forall k \in \mathcal{K} \\
E\{\tilde{r}_k^{R(STAT)}(t_0)|\underline{Q}(t_0)\} - E\{\tilde{r}_k^{B(STAT)}(t_0)|\underline{Q}(t_0)\} \geq \frac{\epsilon}{2}, \forall k \in \mathcal{K}_r
\end{aligned} \quad \text{(C.9)}$$

and therefore, considering (C.7) and the fact that $E\{xy|y\} = yE\{x|y\}$, we have:

$$\Delta_M^{STAT} \leq B - M\epsilon \sum_{l\in\mathcal{L}} \rho_{k_l} Q_{k_l}^{b_l}(t_0) \tag{C.10}$$

For MMW, we can show that similar bound exists with a different constant than B. To see this, first note that by changing the order of summations and stating their bounds based on user sets, $\Phi^X(\underline{Q}(t_0))$ can be written as $\Phi^X(\underline{Q}(t_0)) = \frac{1}{M}\sum_{\tau=t_0}^{t_0+M-1} E\{\Upsilon^X(\underline{Q}(t_0))|\underline{Q}(t_0)\}$ where

$$\Upsilon^X(\underline{Q}(t_0))=\sum_{k\in\mathcal{K}_d}\rho_k Q_k^B(t_0) r_k^{B(X)}(\tau) + \sum_{k\in\mathcal{K}_r}\rho_k [Q_k^B(t_0) - Q_k^R(t_0)] r_k^{B(X)}(\tau) + \sum_{k\in\mathcal{K}_r}\rho_k Q_k^R(t_0) r_k^{R(X)}(\tau)$$

Now, consider a scheduling algorithm called FRM, which maximizes $\Upsilon^X(\underline{Q}(t_0))$ every time slot, i.e., it maximizes a weighted sum of rates for a frame of M-slots, where the weights are based on the queue sizes in the beginning of the frame, i.e., t_0. It is easy to see that $\Phi^{STAT}(\underline{Q}(t_0)) \leq \Phi^{FRM}(\underline{Q}(t_0))$, by using the definition of $\Phi^X(\underline{Q}(t_0))$ as follows:

$$\begin{aligned}
\Phi^{STAT}(\underline{Q}(t_0)) &= \frac{1}{M}\sum_{\tau=t_0}^{t_0+M-1} E\{\Upsilon^{STAT}(\underline{Q}(t_0))|\underline{Q}(t_0)\} \\
&\leq \frac{1}{M}\sum_{\tau=t_0}^{t_0+M-1} E\{\max_X[\Upsilon^X(\underline{Q}(t_0))]|\underline{Q}(t_0)\} \\
&= \frac{1}{M}\sum_{\tau=t_0}^{t_0+M-1} E\{\Upsilon^{FRM}(\underline{Q}(t_0))|\underline{Q}(t_0)\} = \Phi^{FRM}(\underline{Q}(t_0))
\end{aligned} \tag{C.11}$$

where in the inequality we have used the fact that $g \leq \max(g), \forall g$, and the first equality in the last line follows by the definition of the algorithm FRM.

To prove the stability of MMW, we show that $\Phi^{FRM}(\underline{Q}(t_0)) \leq \Phi^{MMW}(\underline{Q}(t_0)) + D$, where D is a constant. Let define $\check{\mathcal{K}}_r$ as the set of indirect users for which $Q_k^R(t_0) \leq \hat{Q}$. For simplicity of the equations, we assume that $\check{\mathcal{K}}_r$ remains unchanged during M slots (Based

on the following discussions, it is easy to see that this only would affect the constant D and does not have any impact on the fact that the drift bound tends to decrease when the queue sizes get large). Then, according to the definition of the MMW weights and considering the fact that MMW maximizes (C.1), for any $\tau, t_0 \leq \tau \leq t_0 + M - 1$, we have:

$$
\begin{aligned}
&\sum_{l\in\mathcal{L}} \rho_{k_l} w_l^{MMW}(\tau) r_l^{MMW}(\tau) = \\
&\quad \sum_{k\in\mathcal{K}_d} \rho_k Q_k^B(\tau) r_k^{B(MMW)}(\tau) + \sum_{k\in\check{\mathcal{K}}_r} \rho_k Q_k^B(\tau) r_k^{B(MMW)}(\tau) \\
&+ \sum_{k\in\mathcal{K}_r - \check{\mathcal{K}}_r} \rho_k (Q_k^B(\tau) - Q_k^R(\tau)) r_k^{B(MMW)}(\tau) + \sum_{k\in\mathcal{K}_r} \rho_k Q_k^R(\tau) r_k^{R(MMW)}(\tau) \qquad \text{(C.12)} \\
&\overset{n_1}{\geq} \sum_{k\in\mathcal{K}_d} \rho_k Q_k^B(\tau) r_k^{B(MMW)}(\tau) + \sum_{k\in\check{\mathcal{K}}_r} \rho_k (Q_k^B(\tau) - Q_k^R(t_0)) r_k^{B(MMW)}(\tau) \\
&+ \sum_{k\in\mathcal{K}_r - \check{\mathcal{K}}_r} \rho_k (Q_k^B(\tau) - Q_k^R(\tau)) r_k^{B(MMW)}(\tau) + \sum_{k\in\mathcal{K}_r} \rho_k Q_k^R(\tau) r_k^{R(MMW)}(\tau) \qquad \text{(C.13)} \\
&\overset{n_2}{\geq} \sum_{k\in\mathcal{K}_d} \rho_k Q_k^B(\tau) r_k^{B(MMW)}(\tau) + \sum_{k\in\check{\mathcal{K}}_r} \rho_k (Q_k^B(\tau) - \hat{Q}) r_k^{B(MMW)}(\tau) \\
&+ \sum_{k\in\mathcal{K}_r - \check{\mathcal{K}}_r} \rho_k (Q_k^B(\tau) - Q_k^R(\tau)) r_k^{B(MMW)}(\tau) + \sum_{k\in\mathcal{K}_r} \rho_k Q_k^R(\tau) r_k^{R(MMW)}(\tau) \qquad \text{(C.14)} \\
&\overset{n_3}{\geq} \sum_{k\in\mathcal{K}_d} \rho_k Q_k^B(\tau) r_k^{B(MMW)}(\tau) + \sum_{k\in\check{\mathcal{K}}_r} \rho_k Q_k^B(\tau) r_k^{B(MMW)}(\tau) - (K - K_1)\hat{Q}\rho^{max} c^{max} \\
&+ \sum_{k\in\mathcal{K}_r - \check{\mathcal{K}}_r} \rho_k (Q_k^B(\tau) - Q_k^R(\tau)) r_k^{B(MMW)}(\tau) + \sum_{k\in\mathcal{K}_r} \rho_k Q_k^R(\tau) r_k^{R(MMW)}(\tau) \qquad \text{(C.15)} \\
&\overset{n_4}{\geq} \sum_{k\in\mathcal{K}_d} \rho_k Q_k^B(\tau) r_k^{B(FRM)}(\tau) + \sum_{k\in\check{\mathcal{K}}_r} \rho_k Q_k^B(\tau) r_k^{B(FRM)}(\tau) - (K - K_1)\hat{Q}\rho^{max} c^{max} \\
&+ \sum_{k\in\mathcal{K}_r - \check{\mathcal{K}}_r} \rho_k (Q_k^B(\tau) - Q_k^R(\tau)) r_k^{B(FRM)}(\tau) + \sum_{k\in\mathcal{K}_r} \rho_k Q_k^R(\tau) r_k^{R(FRM)}(\tau) \qquad \text{(C.16)} \\
&\overset{n_5}{\geq} \sum_{k\in\mathcal{K}_d} \rho_k Q_k^B(\tau) r_k^{B(FRM)}(\tau) + \sum_{k\in\check{\mathcal{K}}_r} \rho_k (Q_k^B(\tau) - Q_k^R(t_0)) r_k^{B(FRM)}(\tau) - (K - K_1)\hat{Q}\rho^{max} c^{max} \\
&+ \sum_{k\in\mathcal{K}_r - \check{\mathcal{K}}_r} \rho_k (Q_k^B(\tau) - Q_k^R(\tau)) r_k^{B(FRM)}(\tau) + \sum_{k\in\mathcal{K}_r} \rho_k Q_k^R(\tau) r_k^{R(FRM)}(\tau) \qquad \text{(C.17)}
\end{aligned}
$$

where the inequalities n_1 and n_5 are obtained due to the fact that for any numbers $y \geq 0$, $x \geq 0$ and $z \geq 0$, we have $yz \geq (y-x)z$. The inequality n_2 is obtained noting that the queue sizes of users $\check{\mathcal{K}}_r$ in the RS are less than $\hat{Q}$ and the inequality n_3 is based on the definition of ρ^{max} and c^{max}. Finally, the inequality n_4 is based on the definition of MMW, as it maximizes (C.1) over all possible scheduling methods, i.e. $\sum_{l\in\mathcal{L}} \rho_{k_l} w_l^{MMW}(t) r_l^{MMW}(t) \geq \sum_{l\in\mathcal{L}} \rho_{k_l} w_l^{MMW}(t) r_l^{FRM}(t)$. Considering equations (C.13) and (C.17), and changing the grouping of the terms, yields:

$$\begin{aligned}
&\sum_{k\in\mathcal{K}} \rho_k Q_k^B(\tau) r_k^{B(MMW)}(\tau) + \sum_{k\in\check{\mathcal{K}}_r} \rho_k [Q_k^R(\tau) r_k^{R(MMW)}(\tau) - Q_k^R(t_0) r_k^{B(MMW)}(\tau)] \\
&+ \sum_{k\in\mathcal{K}_r - \check{\mathcal{K}}_r} \rho_k Q_k^R(\tau)(r_k^{R(MMW)}(\tau) - r_k^{B(MMW)}(\tau)) \\
&\geq \sum_{k\in\mathcal{K}} \rho_k Q_k^B(\tau) r_k^{B(FRM)}(\tau) + \sum_{k\in\check{\mathcal{K}}_r} \rho_k [Q_k^R(\tau) r_k^{R(FRM)}(\tau) - Q_k^R(t_0) r_k^{B(FRM)}(\tau)] \\
&+ \sum_{k\in\mathcal{K}_r - \check{\mathcal{K}}_r} \rho_k Q_k^R(\tau)(r_k^{R(FRM)}(\tau) - r_k^{B(FRM)}(\tau)) - D_1
\end{aligned} \tag{C.18}$$

where $D_1 = (K - K_1)\hat{Q}\rho^{max}c^{max}$. Now, for any queue size Q, we define the maximum change from time slot t_0 to τ as $\delta(\tau) = \max|(Q(\tau) - Q(t_0)| = \max(a^{max}, c^{max})(\tau - t_0)$. Then, we have:

$$\begin{aligned}
&\sum_{k\in\mathcal{K}} \rho_k Q_k^B(t_0) r_k^{B(MMW)}(\tau) + \sum_{k\in\check{\mathcal{K}}_r} \rho_k [Q_k^R(t_0) r_k^{R(MMW)}(\tau) - Q_k^R(t_0) r_k^{B(MMW)}(\tau)] \\
&+ \sum_{k\in\mathcal{K}_r(\tau) - \check{\mathcal{K}}_r} \rho_k Q_k^R(t_0)(r_k^{B(MMW)}(\tau) - r_k^{B(MMW)}(\tau)) + (2K - K_1)\rho^{max}c^{max}\delta(\tau) \\
&\geq \sum_{k\in\mathcal{K}} \rho_k Q_k^B(t_0) r_k^{B(FRM)}(\tau) + \sum_{k\in\check{\mathcal{K}}_r} \rho_k [Q_k^R(t_0) r_k^{R(FRM)}(\tau) - Q_k^R(t_0) r_k^{B(FRM)}(\tau)] \\
&+ \sum_{k\in\mathcal{K}_r - \check{\mathcal{K}}_r} \rho_k Q_k^R(t_0)(r_k^{B(FRM)}(\tau) - r_k^{B(FRM)}(\tau)) - D_1 - (2K - K_1)\rho^{max}c^{max}\delta(\tau)
\end{aligned}$$

where we have used the fact that there are totally $2K - K_1$ number of $Q_k^{B(X)}(\tau)/Q_k^{R(X)}(\tau)$

in each side of the inequality, and have considered the maximum change in the size of all of them from t_0 to τ. Summing over $\tau = t_0, ..., t_0 + M - 1$, taking conditional expectation and noting that $\delta_l(t_0) = 0$ leads to:

$$\Phi^{MMW}(\underline{Q}(t_0)) \geq \Phi^{FRM}(\underline{Q}(t_0)) - D_1 - \frac{2\rho^{max}c^{max}(2K - K_1)\max(a^{max}, c^{max})}{M} \sum_{\tau=t_0+1}^{t_0+M-1} (\tau - t_0)$$

$$\Rightarrow \Phi^{MMW}(\underline{Q}(t_0)) \geq \Phi^{FRM}(\underline{Q}(t_0)) - D \tag{C.19}$$

where $D = D_1 + (M-1)(2K - K_1)\rho^{max}c^{max}\max(a^{max}, c^{max})$.

Based on (C.11) and (C.19), we have $\Phi^{MMW}(\underline{Q}(t_0)) \geq \Phi^{STAT}(\underline{Q}(t_0)) - D$ and as a result, based on (C.7) and (C.10), we have:

$$\Delta_M^{MMW} \leq B + 2DM - M\epsilon \sum_{l \in \mathcal{L}} \rho_{k_l} Q_{k_l}^{b_l}(t_0) \tag{C.20}$$

This bound shows that if the queue sizes tend to grow large, the drift will become negative, which is sufficient to prove the stability of MMW [60].