Scheduling Algorithms for OFDMA Relay Enhanced Cellular Networks

by

Saad Al-Abeedi

A thesis submitted to the Faculty of Graduate and Postdoctoral Affairs in partial fulfillment of the requirements for the degree of

Doctor of Philosophy in Electrical and Computer Engineering

Ottawa-Carleton Institute for Electrical and Computer Engineering (OCIECE)

Department of Systems and Computer Engineering

Carleton University

Ottawa, Ontario, Canada, K1S 5B6

October 2011

Library and Archives Canada

Published Heritage Branch

395 Wellington Street
Ottawa ON K1A 0N4
Canada

Bibliothèque et Archives Canada

Direction du Patrimoine de l'édition

395, rue Wellington
Ottawa ON K1A 0N4
Canada

Your file Votre référence
ISBN: 978-0-494-87753-1

Our file Notre référence
ISBN: 978-0-494-87753-1

The undersigned recommend to
the Faculty of Graduate and Postdoctoral Affairs
acceptance of the thesis

Scheduling Algorithms for OFDMA Relay Enhanced Cellular Networks

submitted by

Saad Al-Abeedi, M.Sc., B.Sc

in partial fulfillment of the requirements for

the degree of Doctor of Philosophy in Electrical and Computer Engineering

Chair, Howard Schwartz, Department of Systems and Computer Engineering

Thesis Supervisor, Roshdy H.M. Hafez

External Examiner, Abraham Fapojuwo, Department of Electrical and Computer Engineering, University of Calgary

Carleton University

October 2011

Abstract

4G wireless networks are developed to support higher throughput demands. Cost effective solutions are being proposed for such networks to enable high speed connectivity with reduced implementation costs. OFDMA combines the benefits of OFDM modulation and the flexibility of two-dimensional system resource allocations. The latest wireless standards are being developed to get the advantage of OFDMA and other new technologies such as Relaying and Adaptive Modulation and Coding (AMC). Implementing AMC is proven to add a substantial gain to system capacity. Relaying is also another technology that can be a cost effective solution to provide extra coverage and/or enhanced capacity. To get the most out of these technologies, efficient scheduling techniques are needed.

In this thesis, we propose new scheduling algorithms for OFDMA based cellular networks adopting Decode-and-Forward relaying. Both centralized and decentralized scheduling techniques are considered. Our aim is to improve resources utilization in order to get better system throughput and enhanced coverage. The Intra-cell Radio Resource Reuse (IRRR) patterns for two and three hop relaying were studied to evaluate spectral efficiency of relay based networks.

Acknowledgments

First of all, praise is due to Almighty ALLAH for giving me the courage to face the complexities of life and complete this Ph.D. successfully.

I am heartily thankful to my parents for their spiritual support and for taking care of me since my birth. The words cannot express my gratitude to them for everything they provided throughout my life.

I wish to express my sincere thanks to my supervisor, Prof. Roshdy Hafez for his support, patience, and motivation. His thoughtful guidance and warm encouragement were of great help in the completion of my Ph.D. work.

I would like also to thank my thesis committee members, Prof. Abraham Fapojuwo, prof. Dimitrios Makrakis, Prof. Mohamed S. El-Tanany and Dr. Len MacEachern for their valuable suggestions and insightful comments.

Finally, I would like to thank my wife Samiyah. Her support and encouragement were the bedrock upon which the past years of my life have been built. Her patience and devotion throughout my years of study and research gave me the strength to successfully overcome many difficulties.

Table of Contents

Abstract iii

Table of Contents v

List of Figures viii

List of Tables x

List of Acronyms xi

Chapter 1
Introduction 1
1.1 Research Motivation 1
1.2 Problem Statement 4
1.3 Contributions 4

Chapter 2
Background and Literature Review 7
2.1 OFDMA based Networks 7
2.1.1 PHY Layer 9
2.2 Relaying 12
2.3 Scheduling 15
2.3.1 Scheduling for single hop networks 17
2.3.2 Scheduling in Multi-hop Relay Networks 19
2.4 Frame Structures For Multi-hop Relaying 31

Chapter 336
Radio Resources Management in Relay Based Cellular Network 36
3.1 Introduction 36
3.2 Radio Resources Reuse and Partitioning for OFDMA Multi-hop Relay Cellular Networks 37
3.2.1 Network layout 41
3.2.2 Co-channel Interference 42
3.2.3 Resources partitioning and reuse schemes 46
3.2.4 Average Effective Spectral Efficiency 50
3.2.5 Power Control 54
3.2.6 Evaluation methodology and channel modeling 55
3.2.7 Results and Discussions 57
3.3 Problem formulation for Multi-hop Scheduling 62
3.2.1 System Description 63
3.3.3 Utility Based Optimization Problem 72

3.4 Proposed Approaches to simplify the optimization problem 73
3.5 Conclusions 75

Chapter 4
Weakest Link Multi-Frame Algorithm 76
4.1 Introduction 76
4.2 Related Work 78
4.3 Contributions 80
4.4 System description 81
4.5 Weakest Link Multi-Frame Forwarding (WLMF) algorithm 82
4.6 Utility Functions 84
4.7 Weakest Link Duo-Frame Forwarding (WLDF) algorithm 88
4.8 Improved Solution 88
4.9 Power Control 89
4.10 Performance Simulation 92
4.10.1 Simulation Models and Parameters 92
4.10.2 Results and discussions: 94
4.11 Conclusions 101

Chapter 5
Tunnel Based Scheduling Algorithm 113
5.1 Introduction 113
5.2 Contributions 116
5.3 System description 116
5.4 Tunneling Concept 121
5.5 Static TBSA Scheduling Algorithm 125
5.6 Adaptive grouping and zone assignment 129
5.7 Link Grouping 135
5.8 Performance Simulation 145
5.8.1 Simulation Models and Parameters 145
5.8.2 Results and Discussions 146
5.9 Conclusions 150

Chapter 6162
De-centralized Scheduling Algorithm 162
6.1 Introduction 162
6.2 Contributions 166
6.3 Distributed Scheduling based on Queue Status (DSQS) 167
6.3.1 System Description 167
6.4 Scheduling based on Queue Status 169
6.4.1 Intra-cell Radio Resources Reuse (IRRR) 171
6.4.2 Allocation at the AN's 175
6.4.3 AE Grouping in DSQS 181

6.5 Distributed Flow Scheduling Based on Generalized Fair Sharing 187
6.5.1 System description 187
6.5.2 Resource allocation of access zone 188
6.5.3 Scheduling in the Relay Zone 192
6.5.4 AE Resource allocation 196
6.5.5 AN Resource allocation 196
6.5.6 Static Zone Allocation 197
6.5.7 Flow scheduling 198
6.6 Performance Simulation 203
6.7 Conclusion 210

Chapter 7
Conclusions and Future Work 223
7.1 Conclusions 223
7.2 Future Work 227
7.2.1 Delay Performance 227
7.2.2 HARQ Support 228
7.2.3 Different Objectives 228
7.2.4 Mobility Support 229
7.2.5 MIMO and advance antenna configurations 231
7.2.6 Cooperative Relaying 231

References 233

List of Figures

Figure 2.1 OFDMA frame structure for WiMAX 12

Figure 2.2 Relay based cellular Network 13

Figure 2.3 Frame structure for WiMAX enhanced by (a) de-centrally controlled or (b) centrally controlled relays [63] 32

Figure 2.4 Operation of the relay zone frame structure proposed in [65] 33

Figure 2.5 Multi-frame Concept 35

Figure 3.1 Deployment for single hop case 47

Figure 3.2 CDF of SIR of all cases 57

Figure 3.3 The average user SIR vs. the distance from the BS 58

Figure 3.4 The scatter plot and fitting curves of average SIR vs. the distance from the BS for cases 1, 4 and 7. 58

Figure 3.5 Average Effective Spectral Efficiency 61

Figure 4.1 Multi-Frame concept for (a) WLMF and (b) WLDF algorithms 102

Figure 4.2 Queue Diagram of WLMF and WLDF system 103

Figure 4.3 Average Cell Throughput vs. users demand with different number of users per cell 104

Figure 4.4 Average Throughput Fairness vs. users demand with different number of users per cell 105

Figure 4.5 CDF of Average Users throughput. 106

Figure 5.1 Schematic Diagram of TBSA Example 123

Figure 5.2 Block Diagram of TBSA Example 123

Figure 5.3 Timing Diagram of TBSA Example 124

Figure 5.4 Cell portioning schemes: (a) three-hop case, and (b) two-hop case. 139

Figure 5.5 Average Cell Throughput vs. users demand with different number of users per cell.....151

Figure 5.6 Average Throughput Fairness vs. users demand of the TBSA system with different number of users per cell.....152

Figure 6.1 Average Cell Throughput of different utility assumption at RS scheduler for three-hop relaying case with 72 UT's with different traffic conditions.....178

Figure 6.2 Cell portioning schemes: (a) three-hop case, and (b) two-hop case.....185

Figure 6.3 An Example of group allocation (a) with no dedicated zone, (b) with a dedicated relay zone.....186

Figure 6.4 Average Cell Throughput vs. users demand for the cell with 72 users, and with different number of hops.....211

Figure 6.5 Average Cell Throughput vs. users demand for the cell with 18 users, and with different number of hops.....212

Figure 6.6 Average Fairness vs. users demand for the cell with 72 users, and with different number of hops.....213

Figure 6.7 Average Fairness vs. users demand for the cell with 18 users, and with different number of hops.....214

Figure 6.8 CDF of Average Users throughput of our algorithms considering two-hop cell layout.....215

List of Tables

Table 3.1 System parameter assumptions 56

Table 4.1 Different utility functions for WLMF and WLDF algorithms 84

Table 5.1 Link Grouping for three-hop case (Figure 5.4-a) 140

Table 5.2 Link grouping for two-hop case (Figure 5.4-b) 141

Table 5.3 Group Resource Allocation Example 141

Table 6.1 Link grouping the two cases shown in Figure 6.2 185

Table 6.2 Different Merits and their objectives 191

List of Acronyms

OFDMA	Orthogonal Frequency Division Multiple Access
3GPP	Third Generation Partnership Project
A&F	Amplify-and-Forward (relaying)
A-DSGFS	Adaptive-DSGFS
AMC	Adaptive Modulation and Coding
AE	Access Entity
AN	Access Node
ARQ	Automatic Repeat Request (an Error control technique)
AS	Access Station (base station or relay station)
ASP	Adjacent Subcarrier Permutation
A-TBSA	Adaptive-TBSA
ATM	Asynchronous Transfer Mode
AWGN	Additive white Gaussian noise
BS	Base Station
CDF	Cumulative Density Function
CID	Connection ID
CNR	Carrier-to-Noise power Ratio
CPF	Centralized Proportional Fair
CS-VF	Centralized Scheduling with Void Filling
D&F	Decode-and-Forward (relaying)
DL	Downlink
DL-MAP	Downlink Mapping
DPF	De-Centralized Proportional Fair
DSGFS	De-centralized Scheduling based on Generalized Fairness
DSL	Digital Subscriber Line
DSP	Distributed Subcarrier Permutation
DSQS	Distributed Scheduling based on Queue Status

E&F	Estimate-and-Forward (relaying)
E2E	End-to-End
FEC	Forward Error Correction (Coding)
FFT	Fast Fourier Transform
FIFO	First-In-First-Out
FR	Full Reuse
HARQ	Hybrid Automatic Repeat reQuest
IEEE	Institute of Electrical and Electronics Engineers
IRRR	Intra-cell Radio Resource Reuse
LOS	Line of Sight
LTE	Long Term Evolution
M&F	Modulate-and-Forward (relaying)
MAC	Media Access Control
MF	Maximum Fairness (Throughput Fairness)
MS	Mobile Station
MT	Maximum Throughput
NCC	Number of Clusters per Cell
NLOS	Non-Line of Sight
OFDM	Orthogonal Frequency Division Multiplexing
OFDMA	Orthogonal Frequency Division Multiple Access
PF	Proportional Fairness
PHY	Physical Layer in the network
PMP	Point-to-Multi Point (networks)
PR	Partial Reuse
R-RTG	Relay Receive/Transmit transition Gap
RS	Relay Station
RSS	Received Signal Strength
SDMA	Space Division Multiple Access
S-DSGFS	Static-DSGFS

SINR	Signal to Interference and Noise Ratio
SIR	Signal to Interference Ratio
SLPE	System Level Performance Evaluation
SNR	Signal-to-Noise Ratio
SS	Subscriber Station
S-TBSA	Static-TBSA
TBSA	Tunnel based Scheduling Algorithm
TCS	Tunnel based Centralized Scheduling
TE	Transmitter Entity
UL	Uplink
UT	User terminal
VoIP	Voice over Internet Protocol
Wi-Fi	Wireless Fidelity
WiMAX	Worldwide Interoperability for Microwave Access
WLDF	Weakest Link Duo-Frame Forwarding
WLMF	Weakest Link Multi-Frame Forwarding

Chapter 1
Introduction

1.1 Research Motivation

In recent years, broadband communications have achieved significant success. Internet over DSL and cable TV are examples of the widely adopted and still growing broadband technologies. However, in rural and certain suburban areas, these technologies face difficulties because of the high infrastructure costs. The latest advancements in telecommunication technologies have opened the door for alternative wireless solutions that are expected to efficiently deliver high data rates to subscribers at a reasonable cost. Moreover, wireless broadband networks can support exciting new mobile applications, making them an attractive solution even at locations where fixed broadband services are well established.

Adaptive Modulation and Coding (AMC) and Orthogonal Frequency Division Multiplexing (OFDM) are two important technologies that have made high data rate communications a reality by virtue of advancements in signal processing and telecommunication theory. With AMC, the system is able to operate at a very high capacity, close to the Shannon limit [1]. In traditional access techniques, such as TDMA and CDMA, time domain equalization is needed to overcome inter-symbol interference (ISI) especially when the system operates at high data rate due to

frequency selectivity of the channel. As the rate increases, time domain equalizers become impractical due to increased complexity. OFDM relaxes the requirement of equalizers, enabling very high data rate communications. Among the several advantages of OFDM, there is one where it can bring a multi-user diversity gain by employing OFDMA. OFDMA features the advantages of OFDM and enhances the performance by, allowing the system to efficiently assign bands, as well as time slots to users based on their channel and traffic conditions. [1]

Wireless networks connect every User Terminal (UT) to the wired telecom infrastructure and its services via gateway nodes. For example, in cellular systems the Base Station (BS) is the UT's gateway to the wired infrastructure, where the link between the UT and the BS is the air. The link between the BSs and the wired infrastructure is usually wired. At times, the cost of installing the BS is high compared to its revenue, especially when the number of covered UT's is too low. We observe this in rural areas, and even at urban locations where a small subscriber base is located in a shadowed region. This situation may preclude service providers from serving those areas. A solution here is to enable multi-hop relaying that allows one or more nodes to relay the traffic of other nodes to/from the BS. Such a node, called a Fixed Relay Station or simply a relay station (RS), may be devoted only for relaying. Another solution is to have the UT's relay each other's traffic. In addition to its benefit in reducing cost of deployment, it may lead to an increase in system capacity, especially in "Manhattan-model" like topologies due to the enhanced Intra-cell Radio Resource Reuse (IRRR) [2],[3].

On the other hand, relaying can lead to problems, such as increased delay due to multi-hopping and reduced overall spectral efficiency in some network topologies, which are mostly seen in open areas with large number of UT's. Relaying will, in fact, increase link capacity due to shorter distances between terminals. However, system spectral efficiency may drop due to an increase in competition on the system's bandwidth caused by relaying. In our work, we are concerned about efficient utilization of system resources by proposing new efficient scheduling algorithms.

Resource scheduling for relay enhanced OFDMA cellular networks has recently gained much interest. The complexity of this problem diverts the literature proposals to enforce limitations for the sake of getting a tractable optimization framework. Ignoring IRRR and restricting the relay system to two hops are examples of such limitations [4]-[11]. It is difficult to find an optimal solution for the scheduling problem involving relay based OFDMA networks with more than two hops [9]. Moreover, the difficulty increases with the adoption of IRRR. Therefore, simplifying the problem may be needed to efficiently find the solution with modest complexity.

Reuse patterns for two hops have been studied in literature, and results have concluded that full reuse can achieve higher system throughput [12]. The reuse patterns for three hop cells have not been considered before. Since one of our goals is to negotiate this limitation, we have included a study for three hop cases to examine the spectral efficiency improvement by utilizing the IRRR capability. We have studied the impact of power control, and AMC mode selection on cell spectral considering one, two and three-hop cases.

1.2 Problem Statement

In order to achieve higher throughput, cellular networks adopting OFDMA are proposed to have a higher IRRR capability. This in turn will make the users at cell edges suffer from low SINR. For this purpose, relays may be used to reduce the capacity outage of those users. Unfortunately, the network consumes bandwidth for the relaying operation, which may reduce the overall throughput of the cell.

We will provide solutions for radio resources scheduling to improve throughput performance by utilizing IRRR and/or by a justifiable simplification of the forwarding mechanism. Both distributed and centralized solutions are considered.

1.3 Contributions

In this thesis, we have studied scheduling and resource allocation problems in the downlink channel of OFDMA relay enhanced cellular networks. Downlink transmission consumes most of the bandwidth of the network as most of the high data rate applications occur in the downlink. Improving IRRR capability is essential in improving the system throughput. The overhead and system complexity precludes the utilization of IRRR. As a result, most of the proposals in literature avoid IRRR in order to get a tractable optimization problem. According to our study of the achievable spectral efficiency of some IRRR patterns for a relay based cellular network, we found that if the reuse capability is not utilized, then there is no need to change the coding and modulation modes on every hop towards the destination. Performance improvement is small compared to the increase in overhead, time and hardware

requirements. We list below our main research contributions generated through this thesis work:

1. A centralized scheduling algorithm referred to as the Weakest Link Multi-frame Forwarding (WLMF) algorithm is proposed. The scheduling framework based on this algorithm enhances system coverage and improves system throughput at low load traffic. Through this algorithm, we assign the weakest link AMC mode to all hops from BS to all relayed users. We, hence, adopt the multi-frame concept [66] to further reduce the system complexity without a noticeable reduction in system performance. Better system performance may be achieved due to the reduction in signaling overhead, delay, and hardware requirements.

2. A centralized scheduling algorithm for tunneled traffic is proposed to utilize IRRR capability and improve system throughput performance. This solution is referred to as Tunnel Based Scheduling Algorithm (TBSA).

3. A new de-centralized scheduling system based on what we refer to as "generalized fairness principle" assigns a desired quality or quantity metric to the traffic at each access relay. The resources are shared based on these credits. This algorithm solution is titled Distributed Scheduling Based on Generalized Fairness (DSGF). DSGF is also able to get the benefit of IRRR to further enhance performance.

4. Another de-centralized fair-scheduling algorithm called De-centralized Scheduling Based on Queue Status (DSQS) was implemented with minimal interaction from the BS. This algorithm has IRRR capability, which helps in enhancing the throughput.

5. Weakest Link Duo-frame Forwarding (WLDF) is an improved scheduling system proposed to allow IRRR by controlling the power level of an interfering node. This system is based on a two-frame structure. While this scheme puts restrictions on the transmitted power at lower hop count RS's, the simplicity, reduced complexity and delay requirements are maintained as similar to the WLMF system; however, with improved system throughput and scalability.

Chapter 2

Background and Literature Review

In this chapter some relevant technologies will be presented. OFDMA based networks are introduced. Then a discussion about multi-hop relaying will be provided. We will provide a literature review for scheduling techniques in single hop OFDMA networks, as well as relay enhanced networks.

2.1 OFDMA based Networks

Orthogonal Frequency-Division Multiple Access (OFDMA) is an access technology based on *Orthogonal Frequency-Division Multiplexing* (OFDM). OFDM has gained a huge success as a digital multicarrier modulation. In multicarrier modulation, the data stream is divided into many parallel data streams with a lower rate. Each lower data stream will modulate a certain subcarrier (tone). These tones are evenly spaced and orthogonal to each other, which provide a spectral efficient solution. Lower data rate means longer symbol duration which tends to make the system operating in frequency flat channel conditions in many environments where the delay spread is being less significant compared to symbol duration, and that will make the induced ISI negligible, relaxing the requirements for complex equalizers. [1]

OFDMA shares with OFDM its benefits, such as

- Reduced transceiver complexity for high data applications, since complex equalizer are not required.
- Robustness against narrowband interference as with coding and interleaving, bits transmitted over a subset tones can be recovered.
- Pilot based channel estimation is simple in an OFDM system compared to single carrier systems. This results in an improved channel quality feedback and better quality coherent demodulation.

Diversity gains can be achieved by enabling OFDMA. Frequency diversity can be realized with the help of coding and interleaving across subcarriers when distributed permutation is used. Distributed permutation is a technique used to randomize the subcarrier allocation.

Improved multiuser diversity can be achieved by OFDMA, too. When the users are assigned a subset of the subcarriers based on their channel quality. More than one user can share the available bandwidth by assigning each user a non-overlapped subset of carriers. It is worth mentioning that overlapping is possible when efficient radio resources scheme, *Space Division Multiple Access* (SDMA), or other advanced antenna systems are used.

Because of its advantages, it became the method of choice by the recent standards and systems. IEEE 806.16 and *Long Term Evolutions* (LTE) have adopted OFDMA as their access technology in their latest releases. Both are competing to gain

success as 4G and beyond networks [13], [1]. While they both have some similarities in their PHY specifications, their upper layers, frame structure, and management system are different. In this work, we will consider the OFDMA PHY specification of IEEE 806.16 [14]. Since IEEE 802.16 and LTE are converging in their system assumptions and specifications especially the ones related to PHY layer, we can say that our work is still applicable to LTE standards.

Worldwide Interoperability for Microwave Access (WiMAX) is a broadband wireless protocol developed for metropolitan area networks. Its system profile specifications are based on the IEEE 802.16 standard. Mobile WiMAX, the latest version based on the IEEE 802.16-e amendment of the standard, became the basis of its future development [1]. In our thesis, the term WiMAX and IEEE 802.16 will be used interchangeably, as they share the same basic specifications. In the following sections we will briefly review some of the specifications concerning the WiMAX physical layer.

2.1.1 PHY Layer

The physical layer of Mobile WiMAX is based on OFDMA. Five classes of bands range from 2.3GHz to 3.8GHz are adopted. In each band class, the common channel bandwidths are 5 MHz, 10 MHz, and 20 MHz with FFT size of 512, 1024 and 2048 respectively, which gives 9.765 KHz bandwidth for each subcarrier. [15]

2.1.1.1 OFDM Symbol structure

There are three types of subcarriers in the OFDM symbols: Data, Pilot, and Null. Data subcarriers carry data. Pilot subcarriers are used for estimation and synchronization while Null subcarriers are used as Guard bands and DC carriers.

The Sub-channelization in Mobile WiMAX supports two types of sub-carrier permutations: *diversity* and *contiguous*. DL PUSC, UL PUSC, DL FUSC are examples of Diversity permutations whereas DL AMC and UL AMC are of contiguous permutation types. A *slot* is the minimum frequency-time resource unit of sub-channelization, which is equal to 48 data tones.

A frame consists of 48 OFDM symbols. 44 Symbols are used for data transmission. Each symbol consists of a number of tones, depending on the bandwidth of the channel. In all cases the tone spacing is 10.94 KHz.

2.1.1.2 Frame Structure

The 802.16e PHY supports TDD, FDD, and Half-Duplex FDD operation; however, in this presentation we will only present TDD. TDD enables adjustment of the downlink/uplink ratio to efficiently support asymmetric downlink/uplink traffic, while with FDD, downlink and uplink always have fixed bandwidths. TDD assures channel reciprocity for better support of link adaptation, MIMO and other closed loop advanced antenna technologies. Unlike FDD, which requires a pair of channels, TDD only requires a single channel for both downlink and uplink, providing greater

flexibility for adaptation to varied global spectrum allocations. Transceiver designs for TDD implementations are less complex and therefore less expensive.

Figure 2.1 illustrates the OFDM frame structure for a Time Division Duplex (TDD) implementation. Each frame is divided into DL and UL sub-frames separated by Transmit/Receive and Receive/Transmit Transition Gaps (TTG and RTG, respectively) to prevent DL and UL transmission collisions. It starts by *Preamble*, used for synchronization, followed by *Frame Control Header (FCH)* to provide the frame configuration information, such as MAP message length and coding scheme and usable sub-channels.

Next, *DL-MAP* and *UL-MAP* messages are included to provide sub-channel allocation and other control information for the DL and UL sub-frames, respectively. The frame includes some control sub-channels, such as *Ranging*, which is used to provide closed-loop time, frequency, and power adjustment, as well as bandwidth requests, *CQICH* for fast feedback channel state information, and *ACK-CH for* HARQ acknowledgement.

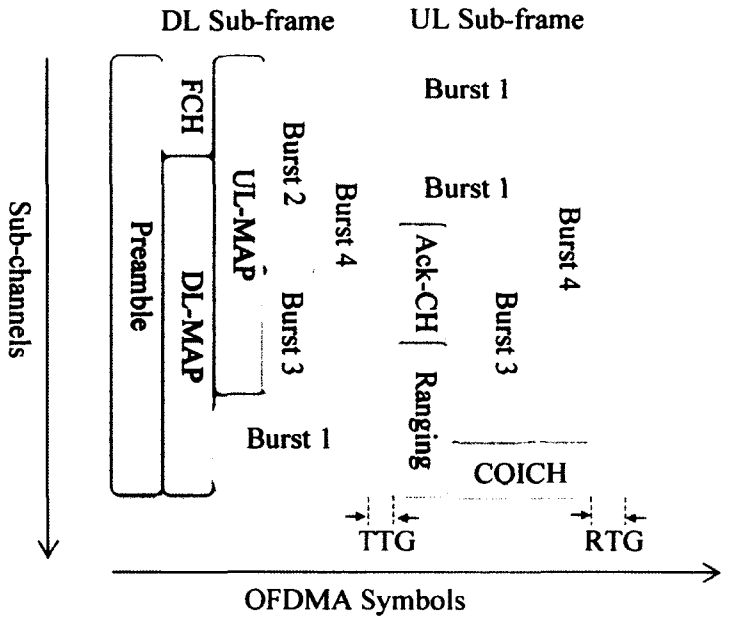


Figure 2.1 OFDMA frame structure for WiMAX

2.2 Relaying

Multi-hop Relaying is a process that retransmits a signal received from a source node and delivers it to the ultimate destination, either directly or through another relay. It acts like a repeater in wired systems. Based on the relaying technique, the signal may be processed for better signal quality. Multi-hopping or relaying mechanisms can be found naturally in different network topologies such as mesh, ad hoc, or sensor networks. Relaying is also featured in infrastructure networks such as WiMAX and LTE, since it offers a cost effective solution to some coverage and capacity problems. Relaying has, therefore, been standardized in IEEE 802.16 and LTE-Advanced.

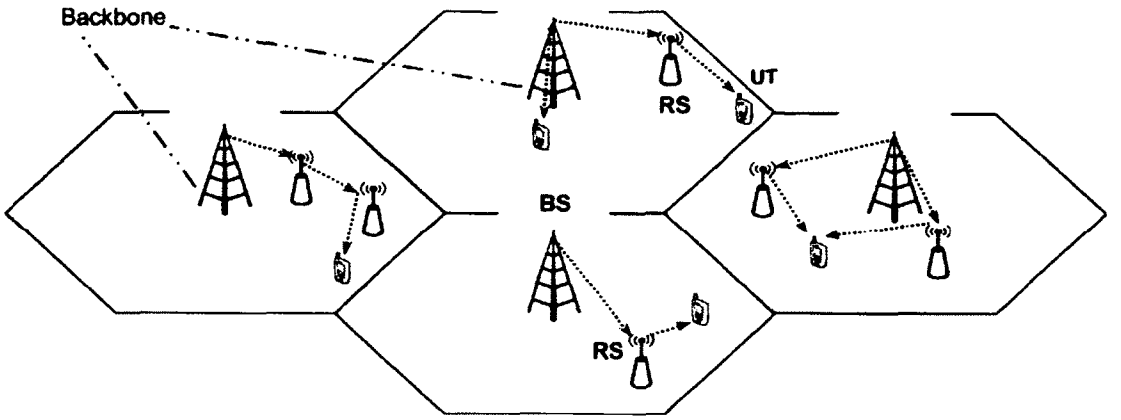


Figure 2.2 Relay based cellular network

In this thesis, we assume a network with a central access node over which all traffic will flow. It is called a base station (BS) and it provides connectivity for all nodes in the network to the wired backbone. It is surrounded by a number of relay station (RS). Each RS is typically owned by the service provider and placed within a specified geographical area to serve a number of UT's. The BS and all RS's serve as access nodes (AN) delivering messages from/to UT's. The RS's could be stationary (fixed) or moving (mobile). In our study we assume fixed RS's only.

Not all RS's are the same. They can be simple and relatively cheap, or may be advanced and expensive, according to their relaying strategy and management options. Management of relay operations can be centralized or distributed. In centralized management, the BS controls the operations and data scheduling of all UT's and RS's. In comparison, distributed management gives the RS the control on some management privileges; more specifically, scheduling and resource allocation. It should be noted that infrastructure networks are centralized by nature. Therefore, some operations still

have to be centrally controlled. For example, the BS will centrally assign resources to each RS, admit UT's, and manages handover between cells.

Relays may cooperate or may not cooperate. In cooperative relaying, more than one AN (one BS and one or more RS's) can cooperate by transmitting replicas of the original signal, which may provide a diversity gain or at least an increased received SINR. Cooperative relaying has received considerable research attention in recent years. Scheduling in the presence of cooperative relaying can be a challenge. Moreover, the diversity gain achieved by cooperation may conflict with the multiuser diversity gain, since a resource that may be used for a user may be consumed by a relay to serve another user for the purpose of cooperation. A balance between the two gains is possible, but only with an increase in complexity. The other option is not to cooperate which simplifies scheduling and enables the potential of utilizing the multi-user diversity capabilities which we intend to pursue. Hence, our proposed algorithms in this thesis do not consider cooperative relaying. In fact, non-cooperative relaying support a simple form of cooperation, mimicking the selection diversity by selecting the best relay based on the throughput or the received SINR.

Relays may differ in their relaying strategies based on how they manipulate the received signal. One strategy is to scale the received signal and retransmit it; this is called amplify-and-forward or A&F in short. This strategy can be viewed as an analog repeater [16], [17]. The relay may demodulate the signal and re-modulate it with the same or different constellation mapping. This strategy is called modulate-and-forward (M&F) [18]. Quantized-and-forward, compress-and-forward, or estimate-and-forward

(E&F) refer to the techniques where an estimate or quantized version of the received signal is forwarded by the relay [19], [20]. The most popular technique is referred to as decode-and-forward (D&F) by which the received signal is decoded, re-encoded, and forwarded to the receiver. However, this method suffers from a larger delay due to the decoding process. Nevertheless, it is still a preferred option due to its higher throughput. Moreover, it is protocol friendly, i.e., it can be applied smoothly with no significant change in network protocols. [21]

IEEE 802.16 standard distinguishes the relay appearance to UT as either transparent or non-transparent. Transparent relay station (T-RS) acts as a distributed antenna for the BS to improve signal quality. In this mode, the UT communicates directly to the BS. In contrast, the Non-Transparent RS (NT-RS) appears as an ordinary BS to the UT, with which it is able to can exchange control messages.

2.3 Scheduling

Scheduling refers to the process by which the UT is allocated its resources, including the UT selection and its resources. Numerous scheduling algorithms have been proposed in the literature for different types of networks, traffic models, MAC, and PHY assumptions. Scheduling the resources can be fixed or adaptive. Assigning fixed resources to all users, regardless of their needs or channel conditions, is simple and fair (in the sense that all UT's have the same resources). However, this is also inefficient, since some resources may not be occupied. Moreover, the occupied resources may not be utilized efficiently. To enhance the utilization, the scheduler is proposed to be queue aware by assigning resources to UT's based on their

requirements. Furthermore, the maximum possible throughput is difficult to achieve without considering channel conditions. For better performance, the scheduler may consider channel conditions, as well as UT requirements.

Opportunistic scheduling is a time varying scheduling technique that is dependent on channel conditions. The main goal of opportunistic scheduling is to enhance resource utilization by allowing the users with a better instantaneous channel quality to transmit, which leads to improved multiuser diversity and enhanced system throughput. This can possibly reduce fairness between users since users in bad channel conditions will never, or rarely, get a chance to acquire resources. A balance between maximizing throughput and fairness is warranted. [7]

One of the key benefits of relaying is to increase the number of higher throughput links. When the scheduler depends on link quality and because we increased the number of flows that have the same improved link spectral efficiency, we expect the fairness will improve as well. Moreover, scheduling using a proportional fair objective function can allow for a good balance between throughput and fairness. It should be noted that relaying without IRRR can have a possible negative impact on the throughput due to increased competition on system resources by relay links at higher demand. In such cases, adopting IRRR can increase the throughput dramatically, but with some degradation in the fairness.

2.3.1 Scheduling for single hop networks

OFDMA scheduling has found a considerable interest in the last few years. Different optimization problems for power, bit, and/or subcarrier allocation with different solutions have been proposed in the literature [22], [23], [24], [25], and [26]. Some solutions attempt to allocate the resources for all demanding users in one block, which may waste some resources if some of those users are in deep fade. On the other hand, if the system allows deeply faded users to wait to give them the chance to have a better channel condition, the system may find a considerable resource utilization improvement as in [26].

Proportional fair scheduler [27] (PF) is a well-known scheduling technique that provides a balance between throughput and fairness. It has been modified to suit different systems, such as CDMA [28], or 1xEV-DO [29]. [30] discussed several PF scheduling methods for multicarrier systems. Modified PF schedulers for OFDMA systems can be found in [22] and [26]. [26] proposed some ways to enhance the performance of the OFDMA scheduler by:

1. Reducing the amount of channel state information (CSI) by selectively and adaptively sending only good frequency clusters, assuming that the whole channel bandwidth is segmented in clusters of the same number of frequency tones.

2. Improving the fairness by using random beam-forming, and in turn producing fast fading rates, increasing the chance to be scheduled and leading to a better fairness and multiuser diversity gain.

Their modified PF serves as a generalized utility based scheduler with a fairness parameter that can tune the scheduler from a maximum sum rate scheduler to a max-min rate scheduler. The two QoS parameters: delay and rate, are taken into account in the scheduler.

In [36], the fairness notion has been taken into account as a weight in the system capacity optimization, where a dual OFDMA optimization of weighted sum ergodic capacity has been formulated for downlink channel of point-to-multipoint (PMP) networks. Each subcarrier is assigned to the highest weighted capacity link user with a power level that is determined by a multi-level water-filling allocation method. Duality theory is used to solve the optimization problem for optimal power and frequency allocation.

Other objectives can be used instead of QoS parameters. For example, in [31] the scheduler was dependent on the business model of an operator network. Customer satisfaction has received an increased attention in research, which may replace the notion of fairness and be the objective function as in [32], [33]. Studies on satisfaction of Internet users showed that if the service quality falls below a certain level, the satisfaction of the user drops significantly. However, if the service level exceeds this threshold, the increase in service level does not necessarily register a significant increase in customer satisfaction level [34], [35]. This behavior is captured by a

logarithmic function which is aligned with the purpose of PF. Thus, PF is considered an attractive solution.

2.3.2 Scheduling in Multi-hop Relay Networks

Multi-hop wireless networks can have a PMP structure, as in sensor networks or employing relays in WiMAX in PMP mode or cellular networks. In this structure, data transmission flows from or to a central node. The other structure can be seen in ad hoc or mesh types of networks, where the communication is point-to-point. In the first case, the downlink scheduling problem is to allocate the resources to the awaiting packets at the central node, link by link, until they reach the ultimate destination, forming a tree type topology which makes centralized scheduling a reasonable option, although others may favor distributed scheduling for scalability, lower delays and faster network response. On the other hand, ad hoc and mesh networks, by nature, require distributed scheduling where many source-destination pairs are contending to get the needed resources.

Scheduling in ad hoc, mesh, or sensor networks has acquired great research interest in the last few years. Many scheduling algorithms have been proposed for different types of MAC, resources, optimization methods, and different goals. Scheduling algorithms designed for such networks [37], [38] is not suitable for multi-hop relaying modes for the lack of a central controller. Furthermore, scheduling algorithms deployed in single hop point-to-multipoint (PMP) systems cannot be used as is, and require modifications that are dependent on the used relaying strategies.

Relays can be used in different ways in conjunction with PMP networks. [39] compares three different architectures that employ multi-hop relaying in cellular networks: Integrated Cellular and Ad hoc Relaying (iCAR) [40], Hybrid Wireless Networks (HWN) [41], and Multi-hop Cellular Networks (MCNs) [42], [43].

In MCNs [42], [43], two types of channels are defined: a control channel of high transmission range used by the BS, and one or more data channels with a lower range. RS's operate in ad hoc mode and are able to route the session locally if the two parties are within the same cell and non-locally through BS, where the transmission is terminated outside the cell. iCAR [40], on the other hand, uses the relays to route the excess traffic from heavily loaded cells to one of the neighboring cells having less load.

HWN has two modes of operation:

1. Cellular mode: for node S to send a packet to node D, the BS decides if the packets will be transmitted through it in a single hop.

2. Ad hoc mode: it will be transferred through other UT's, based on attainable throughput. It is to be noted that the UT is assumed to have ad hoc communication capabilities.

In all cases, it was observed that multi-hop relaying was able to improve the performance compared to regular cellular structure. In best effort service, HWN and MCN were compared. In terms of power consumption, HWN was the best, but it suffered from a significant overhead that was initiated in the ad hoc mode. On the

other hand, MCN offered a higher throughput but mobility significantly degraded the performance compared to HWN. Locality of traffic clearly shows the efficiency of using multi-hopping locally to gain from the limited transmission range and hence the increase in frequency reuse. In real time scenarios, MCN and iCAR have been compared with findings indicating MSC to be better in bandwidth utilization, but becoming less significant with an increase in traffic intensity. This increase in performance comes at the price of high overhead. More descriptions about systems, extensions, and comparisons can be found in [39].

In general, it is possible to classify scheduling systems in relay enhanced networks according to the scheduling of relayed UT flows. If scheduling is done at the BS, it is classified as centralized, otherwise it may be considered distributed, decentralized, or hybrid. However, even the RS with distributed scheduling cannot work without some degree of control by the BS, i.e., the RS operation cannot be fully distributed. In the following section, we review distributed and centralized scheduling systems that are relevant to the scope of this thesis.

2.3.2.1 Centralized Scheduling

The main advantage of centralized scheduling is reducing the cost of RS's due to the simplified RS processing requirements. It further enables a global optimization of network resource assignment. The principal drawback of a centralized system is its slow response to channel events, i.e. the delay in responding to transmission failure is large compared to the distributed case. Furthermore, the validity of channel estimation

is demoted, reducing throughput efficiency due to increased error events and an underestimation of actual channel quality.

In [44] the authors studied the optimal time allocation strategies for a one-dimensional relay enhanced network. They studied how the capacity is affected by changing certain system parameters, such as relay power, relay location, self-noise, and terminal location. They formulated a scheduling problem in a linear program to maximize the total throughput. A sub optimal algorithm was provided in place of solving the linear program.

Scheduling for CDMA networks with relays was studied in [45]. Here, a PF scheduling mechanism was proposed to increase the achievable throughput. Relaying was used to reduce the transmitted power while achieving the required link throughput in CDMA based cellular networks, which in turn reduced the interference to other nodes, thereby increasing the system's throughput. A modified proportional fairness algorithm was proposed to take the effect of the interference from other UT's into account, in addition to the average throughput and instantaneous data rate of the scheduled UT.

A centralized throughput-optimal scheduling based on the instantaneous queue lengths of all access points was implemented in [46] for relay enhanced downlink CDMA networks. All sets of possible active links were first predetermined. Then, at the time of scheduling, the interval of each link was occupied by a source-destination pair optimally chosen according to a utility function that was dependent on the achievable rate and queue lengths.

Another centralized resource scheduling algorithm was introduced in [47] to enable multiuser diversity in TDMA based multi-hop cellular networks. For each user, the best route maximizing the achievable throughput was determined. Then, a scheduling algorithm was applied to optimally choose the best user. Different scheduling strategies were compared, considering the effect of co-channel interference under a multi-cellular environment.

In [48], [49], both cooperative diversity and multiuser diversity were exploited for the same network. Two scheduling algorithms, greedy polling (GP) and partial proportional fair (PPF), were proposed for OFDMA networks with two hops A&F relaying. Data transmission was carried out over a frame of a number of slots and each slot was split into two sub-slots. It was assumed that within one slot, a packet would be transmitted from BS to UT, with or without an RS. When using a RS, two consecutive sub-slots were required, while direct BS to UT transmission required one sub-slot only. The results showed an improved throughput compared to a round robin scheduler under the same network structure and to a PF scheduler in no relay networks.

GP scheduling cyclically polls the users and assigns the resource unit of the highest capacity among the unscheduled units, whereas the PPF uses a two dimensional PF scheduling (in time and frequency), where each resource unit is occupied by the highest scheduling metric. Their results showed that the PPF scheduling algorithm is more suitable for real time services because of its stable

throughput results, while GF is better for the non-real time applications because of its increased system throughput.

In [10], an OFDMA based D&F relay network was considered for scheduling. Scheduling solution start by selecting the optimal route for each flow. Centralized scheduling strategies based on end-to-end channel quality were applied to each flow. Maximum SINR, Proportional fair, and round robin algorithms were modified to serve in the OFDMA environment, which assigned tones to users based on their end-to-end metrics. The same resources assigned for a flow were also assigned to the same flow in all hops. The interaction between multi-hop diversity and multiuser diversity was studied. As expected, their results indicated that the algorithm based maximum SINR would achieve the higher spectral efficiency than a Proportional fair algorithm, and that both were better than round robin. Fairness is not considered in this work.

Through computer simulation, the authors in [50] studied the performance of Wibro (WiMax version in Korea) with relays assuming a round robin scheduler with full buffers. Transparent relays provided throughput enhancement while non-transparent provided coverage extension. Up to 46% downlink improvement and 99% uplink improvement in the throughput of transparent relaying was observed. Coverage extension was achieved by non-transparent relaying, as more than 45% reduction in outage probability was achieved compared to no relay case.

The authors in [51] presented an analytical evaluation to frame efficiency and maximum coverage in two modes of operations of an IEEE 802.16-j system:

1. Single-frame mode: A single frame mode is where backhaul transmissions can be overlapped between relays and or the base station given that a relay cannot receive and transmit at the same time (half duplex assumption). The transmissions from all relays and base stations to all supported relay stations in such cases can be performed in a single frame.

2. Multi-frame mode: The other case assumes no overlapping between backhaul transmissions, requiring multiple frames for the transmission to support relays based on their hop count.

Of these, the single relay mode exhibits better coverage and higher efficiency since it assumes a negligible interference between RS's and BS. It is worth mentioning that the results assume no adaptive modulation and coding (AMC).

In [52], a simulation study was performed for a single hop transparent relay with omni-directional antennae at the RS's. The results showed no significant improvement in the system throughput in downlink cases. It was found that the overhead due to signaling was almost doubled compared to the single hop system. The number and the position of RS's could be optimized for better throughput.

Two allocation methods were proposed in [53] for linear MMR 802.16j networks. Both methods assured throughput fairness: static and dynamic. Both utilized an effective spectrum efficiency index, which reduced the time-consuming dynamic resource allocation. The problem with their proposals was that the bandwidth allocated was based only on channel condition via effective spectrum efficiency, without

considering the dynamic nature of traffic load; i.e. each flow would be assigned resources in a manner inversely proportional to its channel quality to achieve equal throughput among all flows, regardless of traffic conditions.

An adaptive allocation algorithm in [5] was developed to efficiently satisfy TCP bandwidth requirements in a single hop transparent relaying IEEE 802.16j network by balancing uplink with downlink data for both data and ACK packets. For each selected flow (selected based on first come first serve, or FCFS), the resources were time partitioned and assigned to flows in a manner that uplink and downlink TCP traffic were kept balanced. This adaptive algorithm showed noticeable improvement over static systems.

In [6], centralized sum rate maximization was adopted as a resource allocation method in an OFDMA two hop cellular network. An optimization problem was developed to schedule radio resources with and without power allocation. A mixed integer problem was formulated and simplified by relaxing integer constraints to form a linear program that could be solved by standard methods.

The authors in [7] used a graph theoretic approach to jointly optimize routing and subcarrier allocation for centralized two hop relay enhanced OFDMA cellular networks. Their approach maximized the throughput while assuring that a minimum number of subcarriers were allocated to each user.

In [8], the authors relied on heuristics method based on a proportional fair algorithm for a two-hop amplify and forward (AF) relay network. Another two-hop

system was proposed in [54]. A TDM frame partition was assumed where the first was devoted for BS to UT/RS transmission, while the second was for RS to UT. A non-cooperative power allocation game was defined to maximize throughput. An iterative solution was then proposed.

The algorithms presented above suffer from scalability issues, since they are designed for two-hop deployments only. One of the rare proposals that consider more than two-hops can be found in [9], where a centralized scheduling framework was developed. The authors formalized an OFDMA optimal centralized scheduling problem and proofed its NP-hardness and its approximation hardness. Heuristic solutions were then presented. However, their formalization and proposed solutions did not consider the potential of radio resources reuse, and neither did any of the above presented solutions consider this possibility.

To show the potential of radio resources capability, the authors in [55], [56], [44] studied scheduling with reuse capability under Manhattans-like environment where four RS's per cell are placed in spatial isolated locations. The system showed a significant increase in system throughput which can increase further by adopting directional antennas. [2] has employed directional antennas in such environments for OFDMA based networks. Two different strategies have been proposed to schedule RS's transmissions to achieve a high cell throughput assuming RS's, as well as BS, have directional antennas.

2.3.2.2 Distributed Scheduling

Distributed scheduling requires UT flows to be scheduled at the access node (AN). A global optimal solution does not exist because the scheduling problem is broken in pieces where each RS is required to do its optimized scheduling. In addition, the cost of RS's increases due to increase in processing requirements.

The main advantage of distributed scheduling is the ability to respond quickly to error events or any changes in channel conditions. Therefore, the delay encountered in retransmitting erroneous packets, and the errors registered due to imperfect channel estimation are reduced. This results in an improved utilization of system resources and throughput. Utilization is improved because of reduced overheads, which includes those associated with reporting UT flow conditions and receiving management messages, inclusive of the headers of each RS and the addressing information of relayed traffic.

[57] proposed a distributed scheduling algorithm that implemented a scalable factor graph based soft information passing algorithm to schedule flows in TDMA based relay networks. The algorithm determined the valid global collision free schedule based on factor graph modeling and sum product algorithm. The soft passing algorithm was adopted to calculate and transport the soft information that indicated the probability of a link being active.

Distributed resource allocation algorithms were proposed in [58] to enhance both system throughput and coverage of a two-hop OFDMA cellular network. A mix of

FDM and TDM operations was adopted where BS transmission was time-division multiplexed with RS transmission, while FDM was used to separate access transmission from relay transmission. Two resource allocation algorithms based on maximal SINR were proposed; one adopted a fixed TDM between BS and RS transmission, while the other while the other was adapting this TDM operation.

A hybrid scheme has been presented in [59] where radio resources were allocated for both BS and RS transmission based on greedy utility maximization. For a two-hop network and assuming constant power, two sub-frames were allocated; one for BS transmission and the other for RS transmission. The size of each sub-frame was adjusted to accommodate traffic queued at RS; i.e. the resources are allocated to empty RS queues before allocating more resources to users. After RS received the packets, it was required to allocate the assigned resources to the queued traffic. It was also expected to request more resources if there were waiting packets.

A distributed algorithm for a two-hop system was proposed in [60]. Packets were prioritized based on delay and bandwidth requirements and conditions. Policies were developed to create variations in priority levels according to channel status, delay urgency, and fairness conditions.

In [61], a proportional fair allocation algorithm was proposed. This two-hop algorithm tried to perform a joint path selection, power allocation, and sub-channel allocation. Continuous relaxation transformed the problem from mixed integer to linear. Then, a dual decomposition approach is used to solve the optimization problem

in its dual Lagrangeian domain. An attempt was then made to arrive at the optimal solution by an iterative water-filling algorithm.

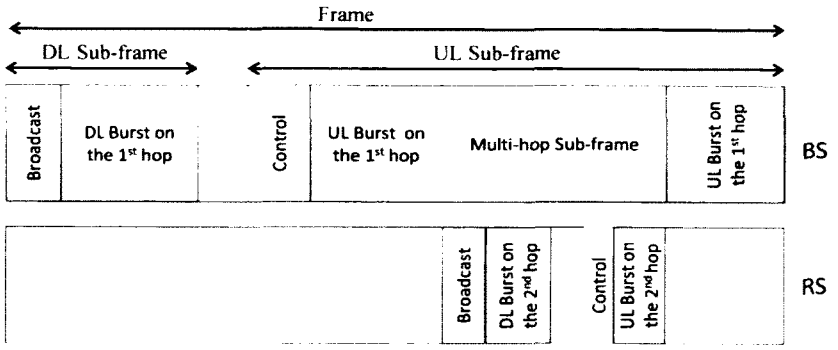
Clearly, the above presented algorithms are restricted to two-hop relay deployment. Moreover, radio resources reuse capability is not considered in these proposals. An attempt to address these issues was presented in [62] where a downlink scheduler was been proposed for IEEE802.16e networks enhanced by fixed relays with directional antennas. First, a routing tree based on hop count was constructed in a manner that enhanced the spatial reuse by assigning routes that did not cause a secondary interference. Secondary interference occurs when two links from different nodes are not spatially independent. Therefore, if routing is unable to eliminate such interference, scheduling is required as a remedy. On the other hand, primary interference happens when the same node engages in two simultaneous communications: in-in, in-out, out-out, which is resolved through scheduling only. The second step was tree-level scheduling where at each time slot, either odd-level nodes or even-level nodes were allowed to transmit, guaranteeing no secondary interference. This simplified the scheduling rule to only addressing primary interference. Then, a proportional fair algorithm was performed at each relay to schedule the relayed traffic as well as the corresponding UT's.

Such a scheme simplifies scheduling while having the advantage of spatial reuse. However, reduced utilization and efficiency are expected due to the endorsed halving of resources. Moreover, combining the routing and scheduling problems leads to an increase in overhead and delays. Typically, fixed relays are well established and

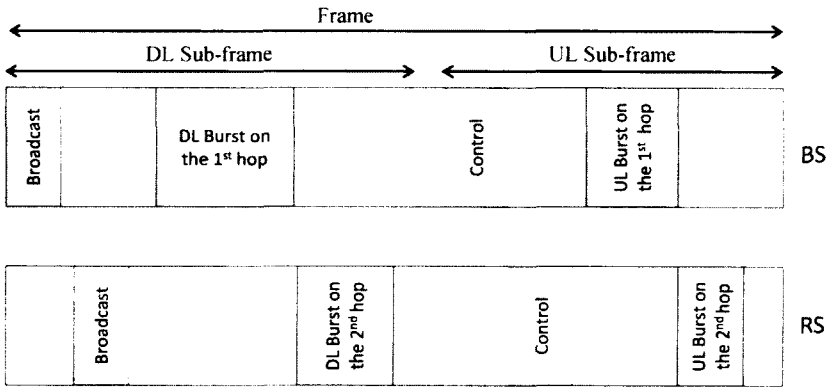
it is unnecessary to change their routing tables with every scheduling instance. The system may require a recovery mechanism to reroute disconnected relays or to enhance load balancing, but this should not be done with every scheduling cycle.

2.4 Frame Structures For Multi-hop Relaying

Understanding the frame structure of multi-hop relay based systems is considered important to a scheduling problem since it significantly affects scheduling algorithms and strategies. In [63], the authors propose two frame structures (Figure 2.3): one for centrally controlled RS's and the other for those controlled de-centrally. For a de-centrally controlled RS (Figure 2.3-a), the BS assigns an interval at the end of the frame for each of its children RS's, whereby they can be used to transmit a sub-frame to their terminals. This sub-frame is seen by terminals as a standard WiMAX frame with a periodic frame header. It follows a D&F relaying principle that processes up to the MAC layer that only manages errors through FEC and ARQ. One transceiver is assumed. Connection masquerading is adopted to manage the data flow connections where the RS acts as BS by which all necessary procedures are handled to setup and manage connections. On the other hand, each RS acts as a regular UT and BS setup, managing the connections to RS only. If another hop is required, another sub-frame has to be allocated at the end of the parent sub-frame. The length of the sub-frame is assumed to be the half of its parent frame or sub-frame.



(a) De-centrally controlled



(b) Centrally controlled

Figure 2.3 Frame structure for WiMAX enhanced by (a) de-centrally controlled or (b) centrally controlled relays [63]

For the centrally controlled relayed system (Figure 2.3-b), the frame starts with the broadcast phase. In this phase, BS broadcast control information includes all management massages for all subsequent hops. Then, a gap is introduced to allow a RS to switch from RX mode to TX, followed by a broadcast of control information to all RS's and UT's that are served by that RS. Data transmission follows the broadcast phase.

It can be observed in both cases that users will see multiple frames since each access point (BS or RS) will transmit its own control messages. One problem in a de-centrally controlled structure is that the frame length is different in the upper level AN from the lower level AN. This may be undesirable in the operation of the system, especially for mobility considerations. Moreover, gaps induced between broadcast messages are a cause of wasted bandwidth.

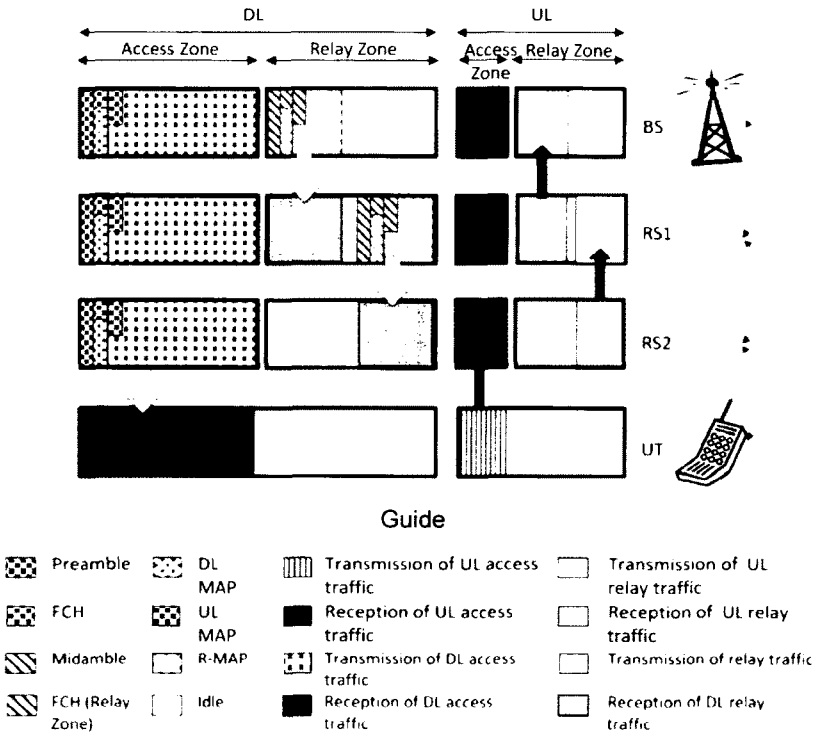


Figure 2.4 Operation of the relay zone frame structure proposed in [65].

[64] and [65] propose a zone based frame structure which divides the frame into different zones for both uplink and downlink. One zone is called the *access zone* where control messages such as FCH and MAP are broadcasted by BS and all RS's. This zone is devoted to communication between AN's and UT's. The other is the *relay zone*, which is associated with communications between relay stations. It should be

Considering UT's scheduling, a competition on resources is observed between first hop communication for UT's served by BS and the i^{th} hop communications for the ones served by the $(i\text{-}1)^{th}$ hop RS's for $i = 2:n$, where n is the maximum hop count. This time span adds a complexity to the scheduling since RS's are also scheduled in relay zones over multiple frames. The question of having the size of the two zones fixed or variable needs to be addressed, since fixing the zones may render some resources unused. On the other hand, making them adaptive increases overhead and can cause stability problems.

A multi-frame structure proposed in [66] has recently been adopted in IEEE802.16-m standard. A number of frames greater than or equal to the maximum hop count in the network were grouped into a single multi-frame. Each RS switched between Subscriber Station (SS) mode and BS mode. In SS mode, the RS acted as a UT; it received from its parent (either RS or BS) control and data messages, and uploaded the scheduled uplink traffic from its subordinate nodes. In BS mode, the RS retransmitted the same control messages (including FCH and MAP), but it only retransmitted data of its subordinate nodes. This structure assumed a centrally controlled RS where the scheduling period was equal to the length of the multi-frame.

Such structures suffer from inefficient resource utilization since the BS would tend to assign resources according the weakest link and will fix MAP entries accordingly. Moreover, because of fixed MAP during a multi-frame, some resources would not be used efficiently.

accordingly. Moreover, because of fixed MAP during a multi-frame, some resources would not be used efficiently.

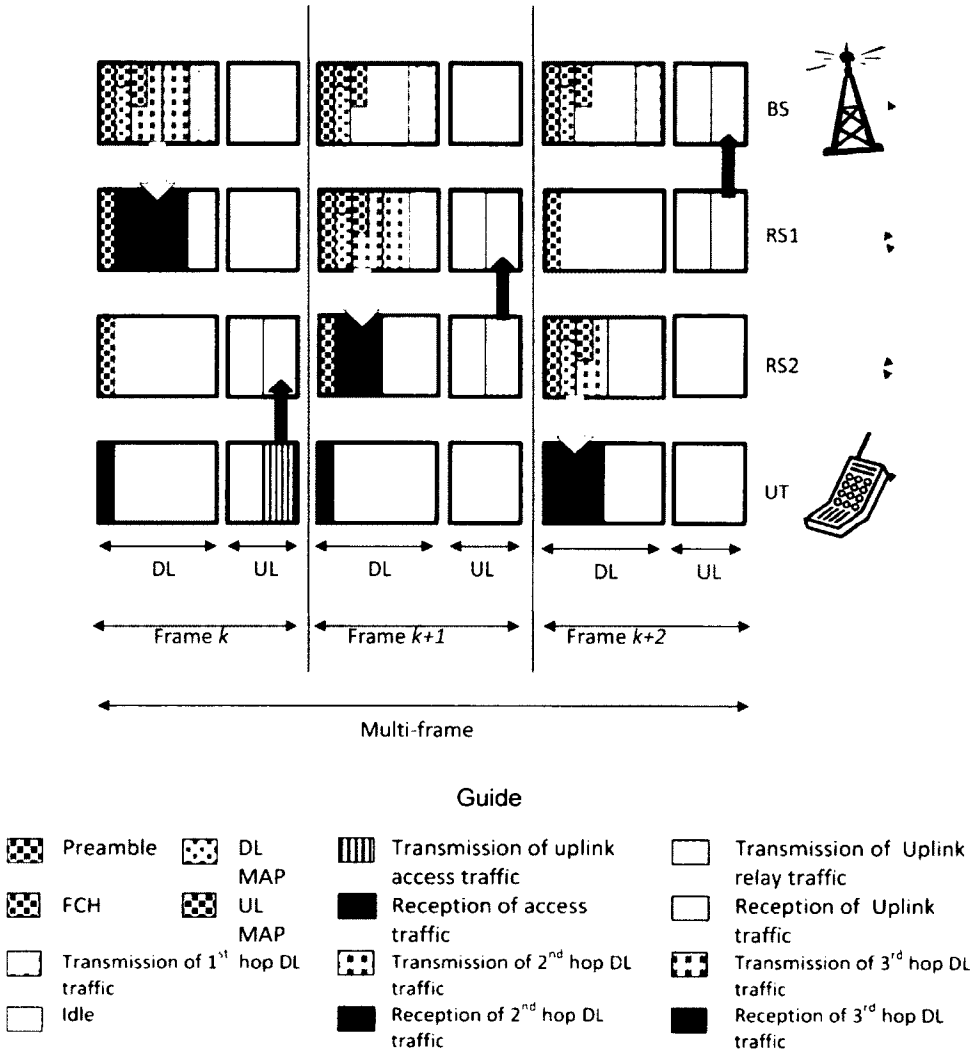


Figure 2.5 Multi-frame Concept

Chapter 3

Radio Resources Management in Relay Based Cellular Network

3.1 Introduction

This chapter will look at Radio Resources Management (RRM) in multi-hop cellular relay networks. RRM features processes and algorithms by which the system controls co-channel interference and reuses radio resources. RRM is classified into two types:

1. Static RRM, which involves cell or network radio planning and includes cellular network structure implementation that involves clustering, sectorization, and partitioning of resources.

2. Dynamic RRM, which dynamically adjusts system parameters and controls the allocation process according to different changes in the system, such as traffic load, number of users, and their channel conditions. User scheduling, radio resource allocation, power allocation, and AMC are examples of dynamic RRM.

Many algorithms, whether static or dynamic, may co-exist in the same network.

This chapter is divided into two parts. The first part considers static RRM schemes related to relay operation, which include radio resource partitioning and reuse patterns. Some reuse patterns are proposed and studied. The second part discusses the dynamic RRM. In particular, we will formulate our optimization problem for optimal user scheduling and radio resource allocation in a multi-hop relay network. We will then provide an overview of our approaches to simplify and solve this optimization problem.

3.2 Radio Resources Reuse and Partitioning for OFDMA Multi-hop Relay Cellular Networks

In this section, we will focus our attention on the effect of IRRR capability, hop count, AMC mode selection, and power control on system throughput performance. Interactions between these factors and their impact on system performance will also be discussed. Achieving improvement from relays is influenced by the ability to reuse resources. This is a critical issue in cellular networks; and in fact, the cellular concept is created based on the radio resources reuse concept.

There is a potential to enhance the coverage with a higher reuse factor using relays [12], [67]-[69]. We can distinguish between two radio resource reuse types: *inter-cell reuse* and *intra-cell reuse*. These are further explained below.

1. Inter-cell reuse is the ability of a cell to reuse the resources available in a network, which is captured by the *Number of Cells in a Cluster, NCC*. Inter-cell reuse involves distributing the radio resources over NCC cells and reusing

all those resources in each cluster. As we reduce NCC, a higher reuse capability is expected due to the reduction of the number of resource partitions at the cost of increased Inter-cell Interference (ICI). The latest standards recommend having NCC equal to one, not only for higher throughput, but also for simplifying radio frequency planning [14].

2. Intra-cell reuse is the ability to reuse the resources within a cell. In this case, partitioning the radio resources involves splitting the cell coverage into sub-cells; each sub-cell may contain one or more relay.

We say that our intra-cell partitioning scheme *fully reuses (FR)* resources if all resources are used by the BS and fully reused by the relay stations. Conversely, if the radio resources are not fully reused, or are fully reused but not utilized fully by BS, we say that the scheme partially reused (PR) the resources. Finally, a no reuse (NR) case is when no resources are reused in the cell. Several resource portioning schemes have been proposed in the literature as part of resource allocation or frequency planning to achieve different goals, such as increasing the coverage or maximizing throughput [12], [67] and [70].

[67] proposed three partitioning strategies where resources are distributed over clusters of one, two, or three cells. No intra-cell resource reuse is allowed in any of these schemes. The effect of a relay channel type, whether it is LOS or NLOS, as well as the location of RS, were studied. The results indicated that a noticeable coverage improvement could be obtained by all three schemes almost equally. However, this gain diminishes in the case of a NLOS relay channel. In [69], different reuse schemes

and path selection rules were considered in the study of resource management policies of two-hop networks. The results indicated that full intra-cell reuse or partial reuse could improve performance. A QoS based path selection algorithm was proposed in [68], [71], which provided improved performance compared to location based algorithms. They employed a full reuse scheme with reuse factor of one which achieved, according to [12], the highest system spectral efficiency compared to the other two-hop schemes.

In [71], another study was presented for resource partitioning and relay location using a simplified channel model. No intra-cell resource reuse was implemented. An approximated 30% reduction in call blocking rate was achieved due to optimal positioning of the relays. [72] studied path selection, as well as a channel ruse algorithm. The path selection at MS link was SIR based, while the RS-BS link was capacity based. Partial reuse scheme is employed in the access zone. Their results showed a throughput enhancement with no significant impact on system coverage. [70] proposed a frequency partition scheme as [73], which does not allow for intra-cell reuse; throughput based relay selection was assumed.

In [14] and [74], a search based algorithm for resources allocation employed a genetic algorithm to jointly optimize path selection, relay location, reuse pattern, and bandwidth allocation to maximize the efficiency of the system. While these solutions may provide decent insights into relay capabilities, they cannot be a practical resource allocation, for the obvious reason of engaging a relay location in the problem.

In [12], four reuse schemes for 2-hop relay systems were studied and the results showed that it was possible to double the spectral efficiency with the appropriate reuse scheme. Their results were based on the omni-directional antennas with simple channel model assumption. However, their results were unrealistic due to the use of a constant capacity formula with no limit on achievable throughput. This resulted in erroneous conclusions.

In our work, we will study the channel reuse problem for 2 hop and 3 hop relay networks, taking into account antenna pattern models, as well as other channel characteristics. Our contributions are summarized as follows:

1. The main objective of this work is to study different partition schemes specifically laid out for two and three-hop relay networks. We will study the impact of radio reuse capability on the cell spectral efficiency. Our study is carried out using some reuse partitioning schemes found in the literature for the two-hop scenarios while for the three-hop cases we proposed *new* radio resources partition schemes.

2. We will study the effect of using the optimal AMC mode on each link and compare this with the case when a lowest link AMC mode is used for all links along the path of each flow. Interestingly, our results show that the impact on cell performance is minimal when no radio resource reuse is implemented, which largely simplifies the forwarding scheme without any significant performance impact.

3. We will study the impact of power control on cell spectral efficiency by implementing simple power control algorithms which were found to be substantial in some cases and insignificant in others. We will discuss the interaction of power control and other factors, namely reuse capability, hop count and AMC mode selection on spectral efficiency.

3.2.1 Network layout

We will estimate the average spectral efficiency for a cell lying in the center of a network and surrounded by 18 interfering cells having the same characteristics such as channel model, relay numbers and location. One, two, and three hop relay deployments are considered. The cell will be sliced depending on the maximum hop count; one, two, or three slices for one, two, and three hops cells, respectively. The cell is sliced in a manner that an equal area and, hence, an equal number of users is assumed in each slice. Uniform user density is assumed in each sector.

Constant power is assumed, and AMC is determined according to the calculated SINR. The AWGN noise is negligible compared to the interference of the surrounding cells. Therefore, we chose to exclude it in our calculations. Two AMC strategies were considered for relaying. The first is called optimized AMC mode selection. This assigns the best AMC mode for each link according to its SINR value. The second is the lowest link AMC, where we assume a single AMC mode on all links over which a certain flow is scheduled.

3.2.2 Co-channel Interference

Inter-cell interference, as well as intra-cell interference, can be experienced by users. If the cell reuses the resources within its coverage, intra-cell interference may impact the throughput of some users. On the other hand, the inter-cell interference results from the surrounding cells utilizing the same resources at the same time. To estimate the cell average spectral efficiency, we calculate the interference experienced by each node, whether the node is a RS or UT, along the path of each flow.

The BS or any RS will be denoted by $a_{j,i}$ where j is the cell index and i is the RS index ($i = 0$ for the BS). For a cell under consideration, indexed by 0, i.e. $j = 0$, the node n will receive from its father node $a_{0,i^{\#}}$. The SINR of slot s at n can be calculated as

$$SINR_n(s) = \frac{P_n(s)}{I_n(s) + N(s)} \tag{3.1}$$

where $P_n(s)$ is the received power, $I_n(s)$ is the interference power and $N(s)$ is the additive noise power experienced by node n at a slot. In an interference limited environment, the additive noise is negligible compared to the interference. Thus it can be ignored and SINR is replaced by Signal to Interference Ratio (SIR). For simplicity, we will drop the slot index, keeping in mind that the calculation is done on slot by slot basis. Thus the SIR at the receiver of node n is given by $SIR_n = P_n/I_n$.

The received power, P_n, is equal to $g_{n \leftarrow a_{0,i^{\#}}} p_{n \leftarrow a_{0,i^{\#}}}$, where $p_{n \leftarrow a_{0,i^{\#}}}$ and $g_{n \leftarrow a_{0,i^{\#}}}$ are the transmitted power and channel gain from $a_{0,i^{\#}}$ to n, respectively. The

interference at node n, I_n, is a combination of inter-cell interference, $I_{n_{inter}}$, and intra-cell interference, $I_{n_{intra}}$, i.e. $I_n = I_{n_{inter}} + I_{n_{intera}}$. The inter-cell interference, $I_{n_{inter}}$, is the interference experienced by the transmission of nodes in the surrounding cells. We will consider the cells in two tiers, thus $I_{n_{inter}}$ can be written in a compact form as

$$I_{n_{inter}} = \sum_{j=1}^{18} \sum_{i=0}^{N_r} y_{j,i,n} I_{a_{j,i} \to n} \quad = Y_n I_{n \leftarrow A} \tag{3.2}$$

where j an index of surrounding cells is, i is an index for relays and base stations where BS takes the value 0. Y_n is an indicator vector containing the elements $\{y_{i,j,n}\}$ given by

$$y_{j,i,n} = \begin{cases} 1 & if \text{ AN } a_{j,i} utilizes \ the \ same \\ & resources \ used \ for \ node \ n \\ 0 & Otherwise \end{cases} \tag{3.3}$$

$I_{n \leftarrow A}$ is an interference power matrix containing the elements $\{I_{n \leftarrow a_{j,i}}\}$ where $I_{n \leftarrow a_{j,i}}$ is the power of the interference from AN $a_{j,i}$. Interference power is calculated based on distance and angle. The angle is used to calculate antenna gain while the distance is needed to calculate the path loss. In general the interference from AS $a_{j,i}$ to node n is given by

$$I_{n \leftarrow a_{j,i}} = g_{n \leftarrow a_{j,i}} p_{a_{j,i}} \tag{3.4}$$

where $g_{n \leftarrow a_{j,i}}$ is the overall channel gain from RS $a_{j,i}$ to node n and $p_{a_{j,i}}$ is the transmitted power by $a_{j,i}$. If we assume that node n is served by a RS indexed by $a_{0,i^{\#}}$, the intra-cell interference can be written as

$$I_{n_{intra}} = \sum_{i=0, i\neq i^{\#}}^{N_r} y_{0,i,n} I_{n\leftarrow a_{0,i}} \quad = Y_{0,n} I_{n\leftarrow A_0} \tag{3.5}$$

$Y_{0,n}$ is an indicator vector that has the elements $y_{0,i,n}$ to show if the user utilized the same resources as $a_{0,i^{\#}}$ or not:

$$y_{0,i,n} = \begin{cases} 1 & if \text{ AN } a_{0,i} utilizes \text{ the same} \\ & resources \text{ used for node } n \\ 0 & Otherwise \end{cases} \tag{3.6}$$

The interference perceived by node n depends on the partitioning associated to the node that is serving it, i.e. node $a_{0,i^{\#}}$. We can replace the index n by $i^{\#}$ in the indicator variables and vectors, which can be rewritten as $y_{0,i,i^{\#}}$ and $Y_{i^{\#}}$, respectively. In addition, the same partitioning pattern is repeated at every cell in every scheme considered here. Thus we can rewrite the indicator variables, $y_{j,i,i^{\#}}$ to be $x_{j,i}\, y_{i,i^{\#}}$, where $x_{j,i}$ is used to indicate if the i^{th} node in the j^{th}cell is transmitting on the designated slot. Notice that $y_{i,i^{\#}}$ doesn't depend on cell index number and is given by

$$y_{i,i^{\#}} = \begin{cases} 1 & if \text{ AN } a_{j,i} \text{ utilizes the same} \\ & resources \text{ as } a_{0,i^{\#}}, \forall j \\ 0 & Otherwise \end{cases} \tag{3.7}$$

and $Y_{i^{\#}}$ becomes a vector of the elements $\{y_{i,i^{\#}}\}$. Thus we can calculate the interference at node n using the following equations

$$I_{n_{inter}} = \sum_{j=1}^{18} \sum_{i=0}^{N_r} x_{j,i} y_{i,i^{\#}} I_{n\leftarrow a_{j,i}} \quad = X_{j,i} Y_{i^{\#}} I_{n\leftarrow A} \tag{3.8}$$

Note that the scheduling matrix, $X_{j,i}$, is of size 18 $\times N_r$, where 18 is the maximum number of surrounding cells while N_r is the maximum number of RS's, which is 0, 6 and 18 for single, two and three hops respectively. The values of $\{x_{j,i}\}$ and hence $X_{j,i}$ are based on scheduler decision on each cell.

Each flow f will be assigned resources according to the links associated with its path. Paths are assumed static and fixed based on relays location. However, access relay selection by node n is not static. Rather, it is dynamic and based on metric maximization. Different metrics may be used for relay selection such as end-to-end spectral efficiency, received signal power or SIR [66], [70].

In our work, maximum SIR is adopted because of its implementation simplicity with relatively good results, especially if the relays are well positioned [70]. Along the path of flow f, different links will experience different interference conditions and have different SIR values and spectral efficiency. To calculate the average spectral efficiency in a cell, we have to estimate the interference for each node along the path of the flow f.

3.2.3 Resources partitioning and reuse schemes

Deployments with one, two, and three hops are considered. The two hop deployments have six relays, while three hop networks assume six single hop relays in addition to twelve 2-hops relays. In each case, the cell is partitioned such that each tear has the same area. Figure 3.1 below shows the reuse patterns under consideration. In this study, we will consider large scale fading only, i.e. path loss and lognormal shadowing. Although small scale fading may negatively impact the spectral efficiency, with strong error control coding, frequency diversity gain of distributed permutation and multi-user diversity gain of channel aware scheduling, the impact on spectral efficiency is reduced. Thus, large scale fading could offer a reasonable estimate of the spectral efficiency as was observed in [12].

In this case radio channels are assumed to be frequency and time invariant. Therefore, time division, (TD) or frequency division (FD) would have the same performance. Therefore, for simplicity, time division (TD) partitioning is assumed. In practice, however, frequency division (FD), or combination of FD and TD can be used, which is governed by the adopted multi-hop frame structure. With FD we may benefit from frequency selectivity of the radio channel to get some frequency diversity gain. Different frame structures can be found in [12], [67] and [70].

In this paragraph, we will briefly describe the partitioning schemes under consideration. These are depicted in Figure 3.1. The partitioning of resources in each case should assure that a *half-duplex* requirement is maintained where the father-node transmission should not interfere with their child nodes' transmission. In addition,

intra-cell interference experienced from surrounding nodes should be tolerable for reliable communication. Note that partitions that have the same color share the resources. However, if the partitions are adjacent to each other, then the sharing is *non-reusable*, which means the resources will be divided between them. On the other hand, if they have the same color, but are not adjacent, they are able to reuse the resources, and hence we they will experience resource *reusable sharing*.

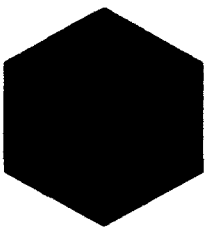
Case Description	**Case Schematic**
Case 1: Single Hop The single-hop case is considered for the purpose of comparison. In this case there is no intra-cell interference since we assume that no reuse is allowed within the cell. For the inter-cell interference, surrounding BS's are the only sources of interference.	

Figure 3.1-a Deployment for single hop case

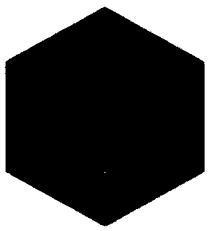
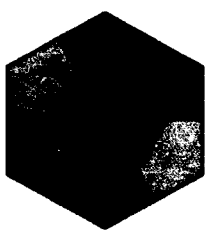
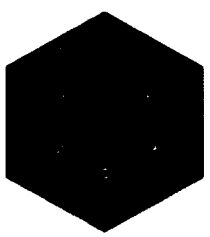
Case 2: 2-Hops No-Reuse (2-Hops NR) [69] presented a 2-hop partition scheme with no IRRR. The resources are partitioned into 7 equal groups, one for the BS and one for each RS. In another deployment proposed in [67] each two adjacent relays dynamically shared the same pool of resources (non-reusable sharing). According to [12], this scheme leads to a small performance loss. In this case, each AN gets one seventh the system resources for both access as well as relay links.	
Case 3: 2-Hops Partial Reuse (2-hops PR) By partial reuse we mean that the resources in this case are reused by RS's only as in [69]. A relay could reuse the same resources of its most distant relay, implying that the pair of RS's would enjoy reusable resource sharing for their access traffic. However, the BS would not be able to reuse the resources [12].	
Case 4: 2-Hop Full Reuse (2-hops FR) In this case, the BS and RS's will be able to reuse the resources. A single sector of the BS shares the same resources as the RS located in the opposite side [12].	

Figure 3.1-b Deployments scenarios for two hop cases

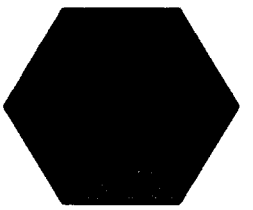
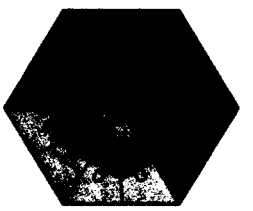
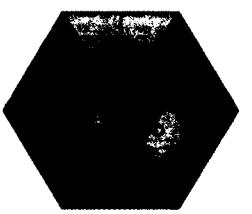
Case Description	Case Schematic
Case 5: 3-Hops No-Reuse (3-hops NR) The resources are partitioned into three sets per sector (assuming a 3 sector cell) with a total of 12 partitions. Within each partition, the adjacent relays share the resources equally. No reuse is assumed in this case.	
Case 6: 3-Hop Partial Reuse (3-hops PR) The resources are partitioned into two sets per sector in a 3 sector cell which results in a total of 6 sets. The BS would reuse the same resources as the 2-hops relays within the same sector. It is assumed that the same resource unit is scheduled for the BS and the third hop link if both belong to the same flow.	
Case 7: 3-Hop Full Reuse (3-hops FR) The resources are partitioned into two sets; each set is used in three sectors in a six sector cell layout. The two sets are used alternatively in each sector. Each set is further segmented into three groups that are used in the manner shown in the figure.	

Figure 3.1-c Deployments scenarios for three hop cases

3.2.4 Average Effective Spectral Efficiency

Effective spectral efficiency is the end-to-end spectral efficiency of a flow that takes into consideration all radio resources used by that flow. It is calculated based on the unit of information, e.g., bits, delivered per radio resources units, e.g., sec × Hz. The efficiency of radio resources consumed by any flow in the cell is limited by the maximum efficiency achieved by a BS link denoted by μ_{BS}. For a SISO system, it is typically equal to 216 bits per slot, which is equivalent to 4.5 bits/s/Hz. If MIMO is used, or a frequency reuse between sectors is allowed, the efficiency is higher. It should be noted that this efficiency does not include MAC layer overhead. Relying solely on the continuous capacity equations as a mean to calculate the spectral efficiency may lead to erroneous conclusions about system spectral efficiency as in [12], which lead to a spectral efficiency more than what is attainable by a SISO WiMAX system. The formula used in [12] appears as follows

$$r = \log_2\left(1 + \frac{-1.5}{\ln(5BER)} SINR\right) \text{bits/sec/Hz} \qquad (3.9)$$

with *BER* equal to 10^{-6} and knowing that the spectral efficiency cannot exceed 4.5 bits/sec/Hz, the maximum SIR supported by this equation is about 22.5dB; the spectral efficiency for more than 22.5dB cannot exceed 4.5 bits/sec/Hz. The SIR for a large percentage of users may exceed this value, and including spectral efficiency with a continuous equation and without clipping could easily yield a wrong conclusion about the achievable average spectral efficiency. It is widely accepted to use a continuous capacity equation as an estimator for system spectral efficiency as in [12], [68], [75],

[76], [72], [70] and [14]. However, without considering the realizable range, erroneous conclusions may be experienced. Thus, proper clipping to the maximum attainable spectral efficiency must be clearly identified; otherwise, the overall average spectral efficiency could be miscalculated as in [12]. Alternatively, we may rely on tables to map SIR to spectral efficiency. This typically furnishes more accurate results.

Our procedure for estimating the average spectral efficiency in a cell starts by defining the *radio resource cost*, which is the number of resource units needed to transmit a single bit. A radio resource unit is a two dimensional unit defined by a time duration and frequency bandwidth (measured by Seconds x Hertz). This cost is the reciprocal of spectral efficiency, μ_l, which can be written for a link l as $1/\mu_l$. For simplicity we will use a slot as our resource unit.

The minimum required radio resources to transmit a single bit for the flow i is $1/\mu_{BS}$. This is the case either when the user is served directly by the BS or when all resources are fully reused in multi-hop cases. In general, we may need extra resources for relaying. For such a case we may write the flow cost, $\mathbb{f}_f$, as

$$\mathbb{f}_f = \sum_{h=1}^{h_f} \mathbb{f}_{f,h} \tag{3.10}$$

where h_f is the maximum hop count for flow f and $\mathbb{f}_{f,h}$ is the cost associated with the h^{th} hop link that belongs to flow f. To calculate $\mathbb{f}_{f,h}$, we need to consider all flows that share the same resources which can be written as

$$\mathbb{f}_{f,h} = \frac{1}{N_v} \max_{\forall v: L^h(v) = L^h(f)} \left(\frac{1}{\mu_{v,h}} \right) \tag{3.11}$$

where $N_{f,h}$ is the total number of flows sharing the same resources as flow f at the h^{th} hop and $L^h(v)$ is the h^{th} link at the route to flow v from the BS; the cost of using the link is divided equally between the flows using the same link.

The cell will be sliced according to the maximum hop count; one, two, or three slices for one, two, and three hops cells, respectively. The cell is sliced such that an equal number of users is assumed in each slice. The way each slice reuses the radio resources greatly impacts the cell spectral efficiency. The most important factor is the ability to reuse the ones at the first hop because it is typically the cell bottleneck through which all traffic passes.

There is no reuse in cases 1, 2 and 5. Therefore, the cost of the flow f is simply

$$\mathbb{f}_f = \sum_{h=1}^{h_f} \frac{1}{\mu_{f,h}}$$

In case 3, the cost of a flow depends on the hop count of the users. The two hop user costs one resource unit for relay links in addition to a half resource unit for an access link since it is shared by another user. Thus flow cost can be written as

$$\mathbb{f}_f = \sum_{h=1}^{h_f} \frac{1 + .5(h-1)}{\mu_{f,h}}$$

In cases 4 and 7, the resources are assumed to be fully reused and the only bottleneck is the BS links to first tier RS's.

$$\phi_f = \frac{1}{\mu_{1,h}}$$

In case 5, the 2 and 3-hops flows will both need only two units. Thus flow cost is given by

$$\phi_f = \begin{cases} \frac{1}{\mu_{1,h}} & for\ h = 1 \\ \frac{1}{\mu_{1,h}} + \frac{1}{\mu_{2,h}} & for\ h > 1 \end{cases} \tag{3.12}$$

The overall flow cost in the cell is $\sum_{f=1}^{N} \phi_f$, where N is the total number of users. The average effective spectral efficiency can be calculated as

$$\bar{\mu} = \frac{S}{B}\left[\frac{1}{N}\sum_{f=1}^{N} \phi_f\right]^{-1} \tag{3.13}$$

where S is the number of available resources units, B is the total bandwidth and N is the total number of flows. This quantity does not represent the actual achievable cell spectral efficiency because the scheduler automatically avoids flows with zero spectral efficiency, which results in a higher average effective spectral efficiency for the cell. Therefore, to get a correct estimation of the average achievable effective spectral efficiency, only schedulable flows are accounted for. Therefore the average spectral efficiency achieved by the cell denoted by $\bar{\mu}_c$ is calculated as follows:

$$\bar{\mu}_c = \frac{S}{B}\left[\frac{1}{N_s}\sum_{f\in G_s} \phi_f\right]^{-1} \tag{3.14}$$

where G_S is a set of size N_S containing the schedulable flows and N_S is the number of schedulable flows. A schedulable flow is identified as the one with SIR exceeds the min SIR.

It is worth mentioning that $\bar{\mu}$ represents the true improvement achieved by relaying, which includes enhancement of outage probability and user spectral efficiency. However, although user efficiency is clearly improved due to the increased link SIR, this is not necessarily enhancing the cell achievable throughput, which is captured by $\bar{\mu}_c$ because of the tradeoff between reducing outage probability and increasing the throughput. The studies in literature [67] - [70] do not distinguish between the two values and use the first one while, in fact, the second one is closely related to what is expected to be achieved in practice.

3.2.5 Power Control

Power adaptation in the downlink is not as critical as the uplink, as the uplink transmission is usually battery operated which makes energy conservation an essential issue. Downlink transmission is typically done by a terminal that is connected to a power source. Nevertheless, power control can potentially reduce interference from surrounding nodes, which can improve the outage performance, increase the throughput and save energy. Many solutions in literature consider power as a resource besides the radio spectrum. [77]

To examine the effect of power control on system performance, we propose to use a simple method that requires no extra overhead or significant complexity. The power control procedure is basically a reduction in power by the transmitting node in a manner that the SIR does not go below the minimum level defined for the assigned AMC mode. With this simple strategy, we anticipate reduced interference, improvement in outage and throughput performance, as well as energy savings.

In Case-6, any third hop link will share the same resources as the first hop serving the same flow. We assume this assignment will be fixed. It can be shown that the SIR values of the two links must be identical to achieve the highest possible throughput. Thus, our power control scheme in this case is to lower the power of the high SIR link in a manner that both interfering nodes receive equal SIR values.

3.2.6 Evaluation methodology and channel modeling

A large number of sampling points (10,000 points) are randomly generated to cover a circular area. At each point, the SIR is estimated where we only consider large scale channel parameters. The points are associated with the AN based on received SIR value. The spectral efficiency is estimated for each point according to the deployment scenario and averaged to get the efficiency of the system.

Moreover, we will assume a NLOS link between the AN (BS or RS) and its subordinate UT's. The BS and RS's will be assumed as Above the Roof Top (ART) with a line-of-sight (LOS) link between each other if they have a direct father-child relationship; otherwise, a NLOS link will be assumed. In LOS cases, we will use the

modified IEEE 802.16 type-D path loss model defined in [14], which assumes a suburban flat terrain environment. The standard deviation of lognormal shadowing is 3.4dB. In the case of NLOS, whether it is between BS/RS and UT or between BS/RS and a subordinate RS, the modified IEEE 802.16 path loss model type B will be used. Such a model represents the intermediate suburban path loss conditions, i.e., moderate tree densities with neither flat nor hilly terrain. The lognormal shadowing of 9.6 dB standard deviation is assumed. [14]

Parameter	**Value**
Frequency band	2.5-2.52 GHz
Bandwidth	20 MHz
Modulation and coding schemes (Convolutional turbo coding) See [78]	QPSK r=1/6,1/3,1/2, 3/4 16 QAM r=1/2, 3/4 64 QAM r=2/3, 3/4
Cell radius	1.5 km
Maximum transmit power BS/RS	43 dBm/40 dBm
Antenna Height of BS/RS/UT	20 m/10 m/ 1.5 m
Antenna gain BS/RS/UT	17dBi/10dBi/0dBi
Antenna model (BS and RS) Note that RS is assumed to have two directional antennas, with $\theta_{3dB} = 70^{o}$, one directed towards the father BS or RS and the other directed to its supported users	$A(\theta) = -\min\left(12\left(\frac{\theta}{\theta_{3\text{dB}}}\right)^{2}, A_m\right)$ dBi $180^{o} < \theta \leq 180^{o}$, $\theta_{3dB} = 70^{o}$, A_m= 20dB for 3 sector antenna
Antenna model (UT)	Omni
UT location	Uniform within the coverage area

Table 3.1 System parameter assumptions

3.2.7 Results and Discussions

The enhancement of SIR as a result of relaying can be observed in Figure 3.2 as we plot the CDF of SIR of each case. This figure shows that increasing the number of hops improves the SIR performance, but this improvement is reduced due to the increased interference caused by intra-cell reuse of the resources. The SIR vs. distance plots are depicted in figures 3.3 and 3.4, which display the improvement of SIR close to cell edges as we increase the maximum hop count. However, this advantage is reduced as we allow IRRR. At 0 dB, we noticed in case 5 that about 2.5% of users are in outage indicating that increased hop count with no IRRR can tremendously enhance the SIR outage performance compared to a single hop case which shows about 35% outage at 0dB. The improvement of SIR outage due to relaying is impacted by IRRR. It was observed in case 7 that full IRRR causes an increased SIR outage of about 10%, whereas partial IRRR caused about 4% in case 6.

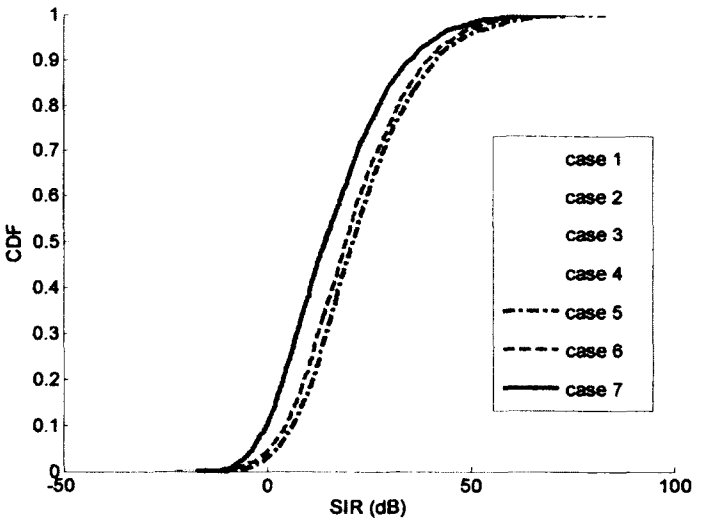


Figure 3.2 CDF of SIR of all cases

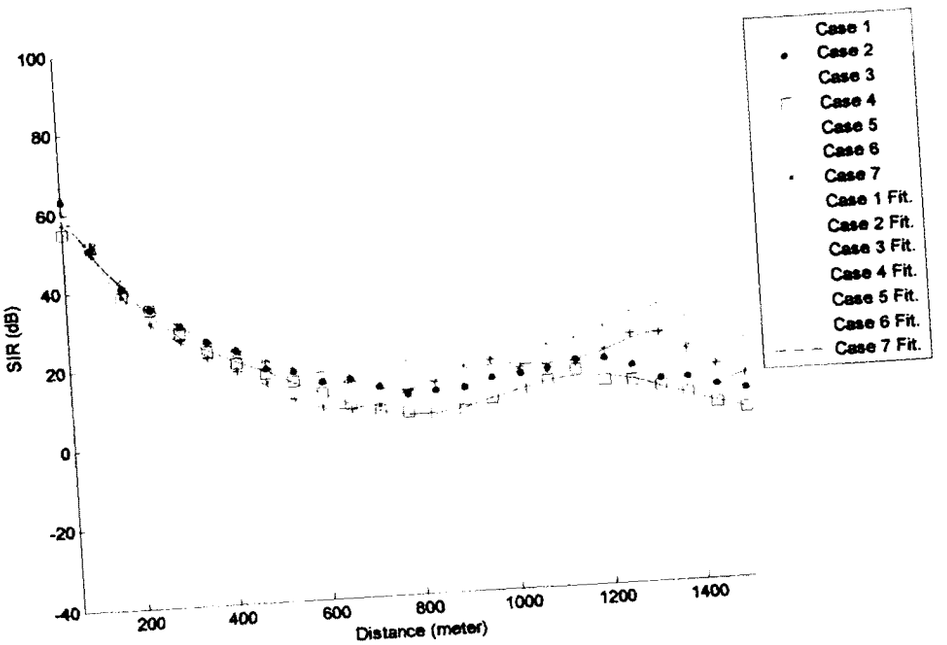


Figure 3.3 The average user SIR vs. the distance from the BS

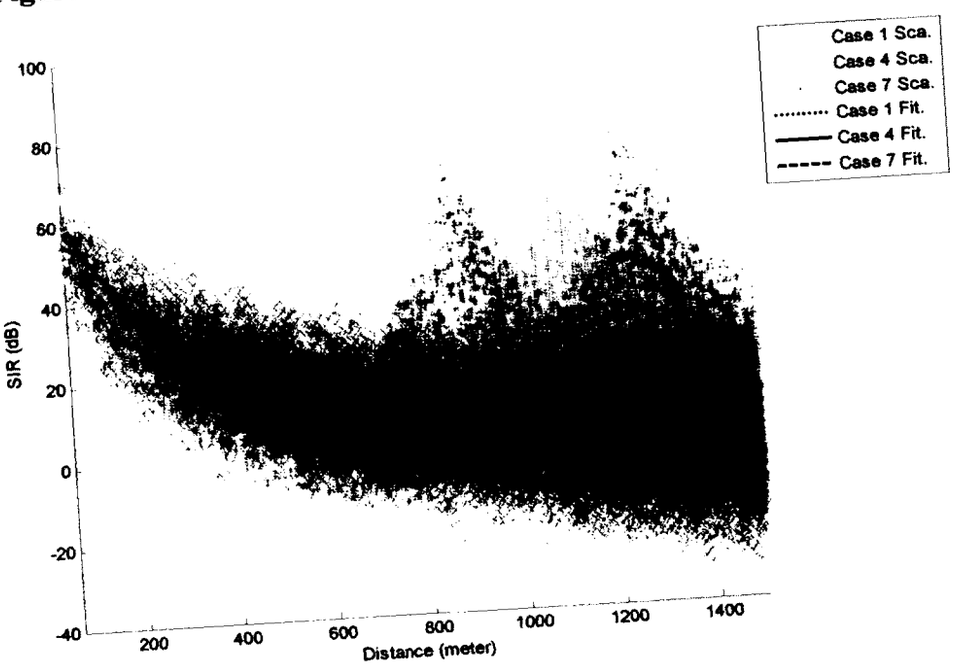


Figure 3.4 The scatter plot and fitting curves of average SIR vs. the distance from the BS for cases 1, 4 and 7.

Figure 3.5 presents the average effective spectral efficiency of each case under different scenarios. Figure 3.5-a shows the results when all flows are considered, while Figure 3.5-b considers the cases when flow with links of SIR<5dB, are excluded. In each sub-plot, we consider the efficiency for each of the seven cases. There are four bars in each case: blue bars are for the cases with no power control, while the red bars are for the cases with power control. The light shaded bars show the cases when all links use the same AMC mode as the worst SIR link, while the darker bars show the improvement from using power control. These bars show the interaction between optimal selection of AMC mode, power control, number of hops, and radio resources reuse.

It is clear from Figure 3.5.a that increasing hop count, in general, increases the average spectral efficiency; this improvement amplifies as we increase the radio resources reuse capabilities, especially in the 3-hop case. Case-6 is an exception; when the lowest link AMC mode selection is adopted, we observe a performance reduction compared to case 5, the no-reuse case, due to the high interference caused by the BS.

Optimal selection of AMC mode did not provide a significant improvement in the cases when radio resources were not reused (cases 2 and 5). Unlike cases when radio resource reuse is adopted, a clear advantage is noticed if the best AMC mode is chosen on each link. The impact of radio resources is apparent with a noticeable increase in the performance, compared to the non-reuse cases.

Figure 3.5.b presents the average spectral efficiency of schedulable users, i.e., we exclude the users that cannot maintain the minimum required average SIR. This

result reflects what we would expect to see from actual system throughput. Different conclusions can be drawn from this figure compared to Figure 3.5.b due to the fact that more users are in outage and cannot be served. This would allow an opportunity for users experiencing good channel conditions to acquire more radio resources. Here, we are assuming full buffer scenarios. The achievable effective spectral efficiency may deteriorate with increasing hop count when we did not utilize the radio resources efficiently. It is evident that full radio resources reuse can improve the spectral efficiency of the system; however, this improvement is noticeably degraded if the optimal AMC mode for each link is not selected. The insignificance of the degradation in the cases when radio resources are not reused suggests that the optimal AMC mode selection may not result in enhanced performance, especially when we consider the overhead, time, and hardware requirements.

Except for case 6 and with the SIR outage flows excluded, the improvement of the average effective spectral efficiency due to power control is minimal at around 2% or less in most cases when optimal AMC mode is used for all links. This improvement is higher when the lowest AMC mode is used for all links, which are slightly increased, especially in case 7 with about 7% improvement. Case 6 is treated differently, where we optimally control the power of the links associated with the three-hop count flows. This case experienced a significant improvement compared to the cases when no power control is adopted. It was able to achieve around 30% improvement for the lowest-link AMC case and around 15% for the optimal AMC case.

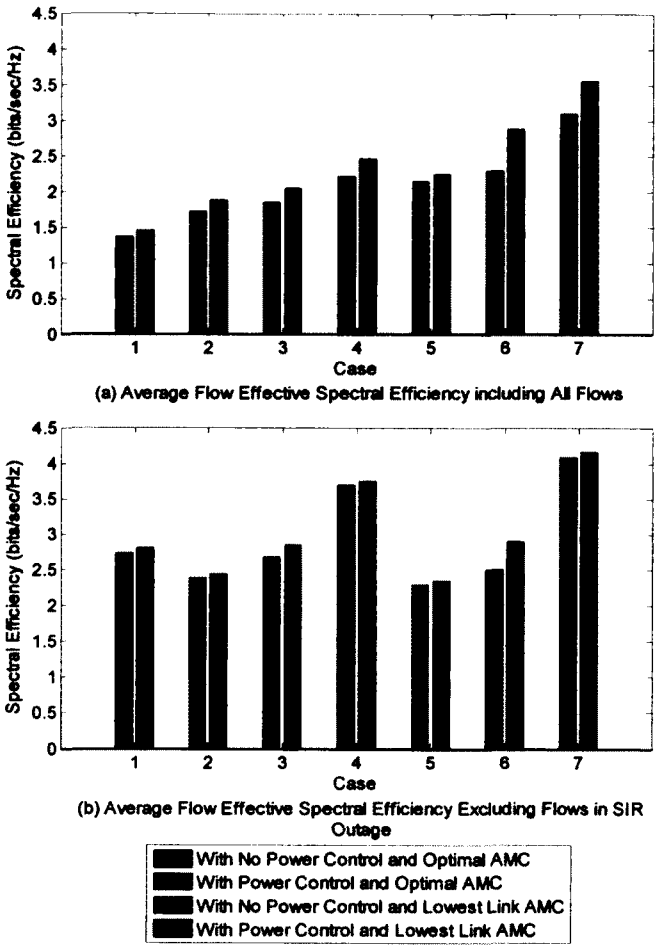


Figure 3.5 Average Effective Spectral Efficiency: subplot (a) shows the cases when we consider the flows that are in SIR outage while (b) shows the cases when the outage flows excluded.

3.3 Problem formulation for Multi-hop Scheduling

Scheduling and resources allocation have been of great interest in recent years. Numerous proposals have been implemented for different system assumptions, objectives, and/or approaches. In the context of multi-hop relay systems, all resource allocation problems involve UT selection and resource allocation for all links, whether they are relay or access links. Furthermore, some solutions have considered queue management or traffic management, and optimal routing or RS selection.

In our formulation, we assume that the routing from the BS to all relays are well established and no further routing change is allowed, though in practice, a rout maintenance mechanism could be added to improve the overall system utilization and reliability. However, this mechanism should not be involved in the scheduling and resources allocation process, as it increases overhead, especially if there are more than two-hop links.

In this section, we will present our problem formulation and present our approaches to solve this problem. We begin by describing our system, followed by an explanation of the constraints involved in this problem.

3.2.1 System Description

We consider a cell in a network being surrounded by 18 cells that interfere with the designated center cell. Each cell is served by a single BS and multiple RS's. UT and RS association with the parent node will be based on the received SINR. This association may be optimized to achieve the highest end-to-end throughput.

Data and management messages will be transmitted over a frame of fixed time duration that comprises two dimensional radio resource units referred to as slots. These slots will be occupied by either Access traffic (A-traffic) or Relay traffic (R-traffic). A-traffic is transmitted by the AN to its local users and carried over AN-UT links while R-traffic carries the traffic intended for its children AN's over what we refer to as AN-AN links. All links are assumed to have a half-duplex operation.

We will use the index j or a to represent AN-AN links, whereas the allocation of AN-UT links will be indexed using the variable i. If we allow $\mathcal{A}$ to be the set of all AN's in the network and $\mathcal{U}(a)$ to be the set of users served by AN a, then we may denote the set of all users in the network by $\mathcal{U}(\mathcal{A})$. We can now define the set of all nodes in the network as $\mathcal{Z} = \mathcal{A} + \mathcal{U}(\mathcal{A})$. Therefore, we have $i \in \mathcal{U}(\mathcal{A}), j, a \in \mathcal{A}$ and $\mathfrak{z} \in \mathcal{Z}$.

A flow is an established path between the BS and a UT, which may pass through one or more RS's. The last RS on the path of a UT is referred to as an Access Node (AN). For the flows associated with AN a, the available slots, $\mathfrak{S}_a$, is defined by a rectangular region specified by a number of frequency and time partitions, $\mathfrak{S}_a^{(f)}$ and

$\mathfrak{S}_a^{(t)}$, respectively. The available resources for RS a are limited by the total number of slots, $\mathfrak{S}$, that is $\mathfrak{S}_a \leq \mathfrak{S}$, $\mathfrak{S}_a^{(f)} \leq \mathfrak{S}^{(f)}$, and $\mathfrak{S}_a^{(t)} \leq \mathfrak{S}^{(t)}$, where $\mathfrak{S}^{(f)}$ and $\mathfrak{S}^{(f)}$ are the total number of frequency and time partitions, respectively. The available slots are indexed by the index s which can take values between 1 to $\mathfrak{S}$ and it is indexed also by f and t, its frequency index (sub-channel or sub-band) and time index (time slot or symbol numbers), respectively: $s = (f, t)$, $s \in \{1, \ldots, \mathfrak{S}\}$, $f \in \{1, \ldots, \mathfrak{S}^{(f)}\}$, $t \in \{1, \ldots, \mathfrak{S}^{(t)}\}$.

The scheduler cannot allocate resources for access traffic at AN a more than $\mathfrak{S}_a$, the total number of slots allocated to AN a. At this point we will assume all AN's can use all the available slots, i.e. $s \in \{1, \ldots, \mathfrak{S}\}$. This implies that AN's can use the same slot simultaneously causing interference to each other. Therefore, the interference from other flows that utilizes the same slot takes an important role in the scheduling problem.

3.2.2 Utility based optimization

All algorithms we propose are based on flow utility maximization. The utility is an objective function defined for each UT to be maximized. Utility functions are devised to give a price for scheduling costs such as radio resources, power and/or other performance measures, such as delay or throughput. Optimizing the utility function is required to achieve the best possible performance according to the desired utility function. Maximum Throughput (MT), Proportional Fairness (PF), and Maximum throughput Fairness (MF) are examples of such utility functions.

Let τ be the time at which the BS transmits a new schedule while $\tau + h_i - 1$ is the time during which flow i receives that schedule. Allowing $u_{i,f,t}(\tau + h_i - 1)$ to be the utility function of flow i on the slot (f,t) scheduled at τ, the following utility functions can be defined at the schedule time τ:

User utility:

$$U_i(\tau + h_i - 1) = \sum_{\forall f \in \mathcal{G}^{(f)}(a)} \sum_{\forall t \in \mathcal{G}^{(t)}(a)} x_{i,f,t}(\tau + h_i - 1)\, u_{i,f,t}(\tau + h_i - 1) \tag{3.15}$$

AN utility:

$$\begin{aligned} U_a(\tau + h_a) &= \sum_{i \in \mathcal{U}(a)} U_i(\tau + h_a) \\ &= \sum_{i \in \mathcal{U}(a)} \sum_{\forall f \in \mathcal{G}^{(f)}(a)} \sum_{\forall t \in \mathcal{G}^{(t)}(a)} x_{i,f,t}(\tau + h_i - 1)\, u_{i,f,t}(\tau + h_i - 1) \end{aligned} \tag{3.16}$$

Cell utility:

$$\begin{aligned} U(\tau) &= \sum_{\forall a \in \mathcal{A}} U_a(\tau + h_a) \\ &= \sum_{\forall a \in \mathcal{A}} \sum_{i \in \mathcal{U}(a)} \sum_{\forall f \in \mathcal{G}^{(f)}(a)} \sum_{\forall t \in \mathcal{G}^{(t)}(a)} x_{i,f,t}(\tau + h_i - 1)\, u_{i,f,t}(\tau + h_i - 1) \end{aligned} \tag{3.17}$$

The scheduler is required to maximize cell utility. However, there are a number of constraints that must be addressed for proper operation of relay. In the following section, we discuss these constraints.

3.2.3 Constraints

An AN may receive R-traffic from its father node. Because of the half-duplex operation of the AN, it should be disallowed transmission while it is listening to its father. This means that the slots used for R-traffic are selected from the available slots competing with A-traffic. We will define an allocation variable $x_{ʒ,f,t}(v)$ which equals 1 when the slot (f,t) is assigned to link ʒ at frame time v and zero otherwise:

$$x_{ʒ,f,t}(v) = \begin{cases} 1 & \text{if link } ʒ \text{ is allocated slot } (f,t) \\ 0 & \text{Otherwise} \end{cases} \tag{3.18}$$

ʒ may represent a AN-AN link or AN-UT link.

The number of slots allocated at scheduling time instance τ for the reception of AN a from its father AN, $\overbrace{x}_a(\tau + h_a)$, is given by

$$\overbrace{x}_a(\tau + h_a) = \sum_{\forall f \in \mathcal{G}^{(f)}(a)} \sum_{\forall t \in \mathcal{G}^{(t)}(a)} x_{a,f,t}(\tau + h_a) \tag{3.19}$$

We used the accent overbrace, $\overbrace{\ }$, to represent the total number or the sum of $\{x_{a,f,t}(\tau + h_a)\}$, where $x_{a,f,t}(v)$ is the allocation variable of the R-traffic received by AN a over slot (f,t).

The AN a will transmit the R-traffic of its children RS's. The RS j is a direct child RS of node a if $j \in D(a)$, where $D(a)$ is the set of the direct child AN's of AN a. Designating $x_{j,f,t}(v)$ be as the allocation variable for the R-traffic received by an AN j over slot (f,t). The number of slots over which an AN j will be able to receive its R-traffic from AN a can be written as

$$\tilde{x}_j(\tau + h_a) = \sum_{\forall f \in \mathfrak{S}^{(f)}(a)} \sum_{\forall t \in \mathfrak{S}^{(t)}(a)} x_{j,f,t}(\tau + h_a) \tag{3.20}$$

$\mathfrak{x}_a^{(R)}(v)$ denotes the total number of slots allocated for whole R-traffic transmitted by AN a, which can be written as

$$\mathfrak{x}_a^{(R)}(\tau + h_a) = \sum_{\forall j \in D(a)} \tilde{x}_j(\tau + h_a) \tag{3.21}$$

If assignment of the slot (f, t) for flow i served by AN a is denoted by $x_{i,f,t}(\tau + h_a)$, then we can write $\tilde{x}_i(\tau + h_a)$, the total number of slots assigned for local access traffic (A-traffic in short) of user i as

$$\tilde{x}_i(\tau + h_a) = \sum_{\forall f \in \mathfrak{S}^{(f)}(a)} \sum_{\forall t \in \mathfrak{S}^{(t)}(a)} x_{i,f,t}(\tau + h_a), i \in \mathcal{U}(a) \tag{3.22}$$

The total A-traffic of AN a will occupy $\mathfrak{x}_a^{(A)}$ slots, given by

$$\mathfrak{x}_a^{(A)}(\tau + h_a) = \sum_{\forall i \in \mathcal{U}(a)} \tilde{x}_i(\tau + h_a) \tag{3.23}$$

Here we use superscript (A) to indicate that the resources are for A-traffic dedicated to users served by AN a defined in the set $\mathcal{U}(a)$.

3.2.3.1 RS Capacity Constraint

The total number of allocated slots cannot exceed the available slot at ANa, $\mathfrak{S}_a$

$$\mathfrak{x}_a^{(A)}(\tau + h_a) + \mathfrak{x}_a^{(R)}(\tau + h_a) \leq \mathfrak{S}_a \tag{3.24}$$

This can be re-written as:

$$\sum_{\forall \mathfrak{z} \in \mathcal{U}(a) \cup D(a)} \sum_{\forall t \in \mathfrak{S}^{(t)}(a)} \sum_{\forall f \in \mathfrak{S}^{(f)}(a)} x_{\mathfrak{z},f,t}(\tau + h_a) \leq \mathfrak{S}_a \tag{3.25}$$

3.2.3.2 Slots Reuse Constraints

A slot can be reused by any node for transmission provided that three conditions have to be satisfied.

1. Output link constraints: no more than one link used for transmission by AN a can be allocated same slot, i.e.

$$\sum_{\forall j \in D(a)} x_{j,f,t}(\tau + h_a) + \sum_{\forall i \in \mathcal{U}(a)} x_{i,f,t}(\tau + h_a) \leq 1 \tag{3.26}$$

   This is equivalent to:

$$\sum_{\forall \mathfrak{z} \in \mathcal{U}(a) \cup D(a)} x_{\mathfrak{z},f,t}(\tau + h_a) \leq 1 \tag{3.27}$$

2. Interference constraints: A slot cannot be reused by a node if it interferes with another node that occupies that slot. We refer to this node as $\mathfrak{z}'$ and we assume an interference group for each node, $\mathcal{I}(\mathfrak{z})$, has been set according to network conditions. It is assumed that an interference set $\mathcal{I}(\mathfrak{z})$ is formed for each node $\mathfrak{z}$. This set can be determined by system design or by node $\mathfrak{z}$ itself, which can measure and determine the main interferers and forward the results to the BS. No transmission is allowed if interference is experienced with node $\mathfrak{z}$ reception:

$$x_{\mathfrak{z},f,t}(\tau_x) + \sum_{\forall \mathfrak{z}' \in \mathcal{U}(\mathcal{J}(\mathfrak{z})) \cup \mathrm{D}(\mathcal{J}(\mathfrak{z}))} x_{\mathfrak{z}',f,t}(\tau_x) \leq 1 \tag{3.28}$$

where $\tau_x \in \{\tau - h_{max} + 1, \dots, \tau, \dots, \tau + h_{max} - 1\}$ is a time instance during which the scheduler may cause interference between different AN's.

3. Half-duplex constraints: an AN cannot receive and transmit at the same time, i.e., we are assuming half duplex relays. Therefore, we need a time division between reception and transmission, which is represented by the following relation

$$\mathbb{I}_{a,t}(\tau_x) + \mathbb{I}_{a^+,t}(\tau_x) \leq 1, \qquad \forall a^+ \in \mathcal{U}(a) \cup \mathrm{D}(a) \cup \mathfrak{F}(a) \tag{3.29}$$

where $\mathbb{I}_{a,t}(\tau_x)$ and $\mathbb{I}_{a^+,t}(\tau_x)$ are indicator functions denoting if an allocation slot or more during time index t is allocated for receiving traffic by AN a and AN/UT a^+, respectively. These are given by:

$$\mathbb{I}_{a,t}(\tau_x) = x_{a,1,t}(\tau_x) \cup x_{a,2,t}(\tau_x) \cup \dots x_{a,\mathfrak{S}(f),t}(\tau_x) \tag{3.30}$$

$$\mathbb{I}_{a^+,t}(\tau_x) = x_{a^+,1,t}(\tau_x) \cup x_{a^+,2,t}(\tau_x) \cup \dots x_{a^+,\mathfrak{S}(f),t}(\tau_x) \tag{3.31}$$

Here, $a^+$is a node that cannot receive at the same time as AN a, which is the case when a^+, one of the local users, $\mathcal{U}(a)$, one of the direct children AN's, $\mathrm{D}(a)$, or the father AN, $\mathfrak{F}(a)$.

3.2.3.3 AN Throughput Constraints:

The total data carried by AN a, $R_a(v)$, should accommodate the local traffic of all of its children RS's, in addition to its local transmission,

$$R_a(\tau + h_a) \geq \sum_{\forall j \in a \cup \mathfrak{C}(a)} T_j(\tau + h_j) \tag{3.32}$$

where $\mathfrak{C}(a)$ is the set of all children RS's served by AN a , h_j and h_a are the hop counts of RS j and a, respectively, and $T_j(v)$ is the total user throughput at AN j. $T_j(v)$ represents the local traffic of AN j, which is a function of the resources allocated to its local access users, as well as their link qualities. It can be expressed as:

$$T_j(\tau + h_j) = \sum_{i \in \mathcal{U}(j)} \sum_{\forall f \in \mathfrak{S}^{(f)}(j)} \sum_{\forall t \in \mathfrak{S}^{(t)}(j)} x_{i,f,t}(\tau + h_j)\, T_{i,f,t}(\tau + h_j) \tag{3.33}$$

where $T_{i,f,t}(\tau + h_j)$ is the throughput of user i at slot (f, t). On the other hand, the total data transmitted by AN a, $R_a(v)$, can be written as

$$R_a(\tau + h_a) = \sum_{\forall \mathfrak{z} \in \mathcal{U}(a) \cup D(a)} \sum_{\forall f \in \mathfrak{S}^{(f)}(a)} \sum_{\forall t \in \mathfrak{S}^{(t)}(a)} x_{\mathfrak{z},f,t}(\tau + h_a)\, T_{\mathfrak{z},f,t}(\tau + h_a) \tag{3.34}$$

where $T_{\mathfrak{z},f,t}(\tau + h_a)$ is the throughput of link a- $\mathfrak{z}$ at slot (f, t) determined according to the received $SINR_{\mathfrak{z},f,t}(\tau - h_{\mathfrak{z}})$.

Other constraints can be considered.

3.2.3.4 Power Constraint:

We may assume that the transmitted power of AN a, P_a, cannot be exceeded. Hence,

$$\sum_{\forall j \in D(a)} \sum_{\forall f \in \mathfrak{S}^{(f)}(a)} P_{j,f,t}(\tau - h_{\mathfrak{z}}) + \sum_{\forall i \in \mathcal{U}(a)} \sum_{\forall f \in \mathfrak{S}^{(f)}(a)} P_{i,f,t}(\tau - h_{\mathfrak{z}}) \leq P_a \tag{3.35}$$

where $P_{j,f,t}(v)$ and $P_{i,f,t}(v)$ are the transmitted power by AN a at slot (f,t) for traffic to AN j and user i, respectively. The above constraint can be written in a compact form as

$$\sum_{\forall \mathfrak{z} \in \mathcal{U}(a) \cup \mathrm{D}(a)} \sum_{\forall f \in \mathfrak{S}(f)(a)} P_{\mathfrak{z},f,t}(\tau - h_{\mathfrak{z}}) \leq P_a \tag{3.36}$$

To simplify our formulation, we will exclude the power constraint and assume a fixed power level. For a single-hop network with AMC, it has been shown that power control in the downlink is of less significance than in the uplink [77]. For the multi-hop case, especially with IRRR capability, power control may help in improving the performance as we saw in the previous section. However, multi-hop relaying suffers from feedback inaccuracy and increased overhead, which render the centralized power control to be impractical. [77]

3.3.3 Utility Based Optimization Problem

As we mentioned earlier, resources (allocation slots and transmitted power) are allocated such that the desired objective function, $U(\tau)$, is maximized. In this case, we wish to find $\{x_{\mathfrak{z},f,t}(\tau_x^+)\}$ that maximize $U(\tau)$, where $\tau_x^+ \in \{\tau, \tau+1, \dots, \tau+h_{max}-1\}$. Let $\boldsymbol{X}$ be 4-D matrices of size $\mathcal{Z} \times \mathfrak{S}^{(f\_max)} \times \mathfrak{S}^{(t\_max)} \times h_{max}$ containing the variables $\{x_{\mathfrak{z},f,t}(\tau_x^+)\}$, where *f_max* and *t_max* are the maximum number of frequency and time partitions, respectively. The optimization problem can be written as

$$\max_{\boldsymbol{X}} U(\tau) \tag{3.37}$$

Subject to

$$\begin{aligned}
&\text{C1: } T_i(\tau + h_i - 1) \le b_i(\tau), \qquad \forall i \in \mathcal{U}(\mathcal{A}) \\
&\text{C2: } x_{\mathfrak{z},f,t}(\tau_x) + \sum_{\forall \mathfrak{z}' \in \mathcal{U}(\mathcal{I}(\mathfrak{z})) \cup \mathrm{D}(\mathcal{I}(\mathfrak{z}))} x_{\mathfrak{z}',f,t}(\tau_x) \le 1, \quad \forall \tau_x \in \{\tau - h_{max} + 1, \dots, \tau, \dots, \tau + h_{max} - 1\} \\
&\forall RS\ a: \\
&\text{C3: } \mathbb{I}_{a,t}(\tau_x) + \mathbb{I}_{\mathfrak{z},t}(\tau_x) \le 1, \qquad \forall \mathfrak{z} \in \mathcal{U}(a) \cup \mathrm{D}(a) \\
&\text{C4: } \sum_{\forall \mathfrak{z} \in \mathcal{U}(a) \cup \mathrm{D}(a)} \sum_{\forall t \in \mathfrak{S}^{(t)}(a)} \sum_{\forall f \in \mathfrak{S}^{(f)}(a)} x_{\mathfrak{z},f,t}(\tau + h_{\mathfrak{z}}) \le \mathfrak{S} \\
&\text{C5: } \sum_{\forall \mathfrak{z} \in \mathcal{U}(a) \cup \mathrm{D}(a)} x_{\mathfrak{z},f,t}(\tau + h_{\mathfrak{z}}) \le 1 \\
&\text{C6: } \sum_{\forall \mathrm{j} \in a \cup \mathfrak{C}(a)} T_j(\tau + h_j) \le R_a(\tau + h_a) \\
&x_{\mathfrak{z},f,t}(v) \in \{0,1\}, \mathfrak{z} \in \mathcal{Z}, f \in \mathfrak{S}^{(f)}, t \in \mathfrak{S}^{(t)}
\end{aligned}$$

This problem has a complex form of generalized assignment problem (GAP) that is known to be NP hard, which means that the solutions that we get from search

algorithms cannot be accomplished in polynomial time. Thus, we will seek some simplifications to efficiently solve the problem.

3.4 Proposed Approaches to simplify the optimization problem

We will adopt three different approaches to tackle this problem:

1. No IRRR:

   In this approach no IRRR is implemented and the number of transmitting nodes to only one. Here we adopt the multi-frame concept presented in section 2.4. The Multi-frame concept has been proposed in [79] and has also been adopted in the IEEE802.16-m standard. The idea of multi-frame is to extend the scheduling interval to a time duration of a number of frames equals the maximum hop count and restrict the activated relay link to the frame number that is equal to the hop count of that link. As an example, if the maximum hop count is 3, then the scheduling interval should be three frames. The first link is for the BS to transmit all relay traffic, the second frame is for the one-hop RS's to transmit their traffic, and the third frame is devoted for the access traffic of 3-hop UT. By this procedure we satisfy the second, third and fourth constraints in our scheduling problem.

   We propose a scheduling algorithm named "Weakest Link Multi-Frame" (WLMF) which takes this approach and further simplifies the procedure by fixing the AMC mode and resource allocation on all links on each frame

within the multi-frame. This will satisfy the sixth constraint. We will discuss this algorithm in the next chapter.

2. Partial IRRR:

   The second approach is a modified version of the WLMF algorithm. In this new proposed scheme, we allow only two frames in a multi-frame, and force the relay links to be reused simultaneously on every second hop. This scheme uses power control, and by fixing the AMC mode and resources allocation as in WLMF, we can satisfy constraints 2, 3, 4, and 6. This approach is referred to as Weakest Link Double Frame (WLDF). This algorithm is also discussed in the next chapter.

3. Full IRRR:

   The third approach divides all links into groups, such that the links in each group can be activated at the same time allowing full IRRR. The links are grouped such that constraints 2, 3, and 4 are met. Three proposed algorithms use this approach:

   a. Tunnel Based Scheduling Algorithm (TBSA)

   b. Distributed Scheduler based on Queue Status (DSQS)

   c. Distributed Scheduler based on Generalized Fair Sharing (DSGFS)

   While grouping the links could be static, the resources allocated to each group can be dynamically tuned in response to changes in traffic and channel conditions. In the following chapters we will study these algorithms.

3.5 Conclusions

In this chapter, we have studied the impact of IRRR, number of hops, power control and the selection of optimal AMC mode on the spectral efficiency considering different radio resources partitioning and reuse patterns for two and three-hop cellular relay networks. Power control had registered some improvement to spectral efficiency. This improvement is more significant in case 6, where the BS uses the same resources scheduled for the third hop links. We further deduced that IRRR is essential to achieve the highest possible capacity from relay enhanced network. In this case, optimizing AMC choice is essential to achieve the desired performance enhancement. If we sacrifice the reuse capability in order to improve the SIR coverage and to simplify the forwarding algorithm, our results show that by choosing the worst link AMC mode for all links, there will be no significant impact on performance. In fact, the reduced overhead, time and processing requirements may lead to better performance.

In the next section, we formulate the scheduling problem for multi-hop OFDMA cellular networks and provide our approaches to overcome its complexity. We have proposed to use the multi-frame concept as a method to simplify and solve the allocation problem without IRRR or with partial IRRR. To support full IRRR, we have proposed a relay link grouping as a further solution for the scheduling problem.

Chapter 4

Weakest Link Multi-Frame Algorithm

4.1 Introduction

Due to their benefits, relays have been adopted in recent standards, such as IEEE 802.16-j, m and LTE-Advanced. Networks enhanced by relays are of much interest in research due to their ability to provide a cost effective solution for improving coverage without an impact on throughput [80]. Such improved performance is a direct result from close proximity of RS to UT's, providing better channel conditions.

However, this performance enhancement comes at the price of increased hardware and overhead cost [4]. There are additional requirements to the system such as optimizing relay locations, routing algorithms to connect them with the BS, scheduling, and resource allocation for both the BS and RS's. These aspects and more are currently active topics in research.

Location of the relays can be optimized to maximize system spectral efficiency or to maximize system throughput with the constraint of specific user mean rate as part of frequency planning [67], [70]. In practice, relay location has to be optimized based on the topology and users' distribution, which is beyond the scope of this thesis.

Routing, on the other hand, is the process to find the path from the BS to the UT. Unlike relay selection where the route from the BS to the access RS is assumed to be fixed and the UT is associated to a RS based on a certain metric, like the received signal power, the SINR, the end-to-end spectral efficiency or the throughput as studied in [73] and [74]. According to [74], when the relay is well positioned, the performance improvement of end-to-end spectral efficiency based solution is not significant when compared to the SINR, but the improvement in efficiency over power based is still significant. For the two-hop cells, routing is the same as relay selection. Routing can be part of scheduling problem as in [77], where a joint routing and scheduling were considered.

Relay implementation adds to the complexity of the allocation problem since the dimensionality of the problem increases. Allocation is carried out over all hops, all time slots and all frequency sub-channels for different user links, as well as relay links. As hop count increases, the overhead increases, in addition to the complexity of the algorithm. Therefore, implementing an efficient and low overhead algorithm is wanted.

4.2 Related Work

Many scheduling algorithms have been proposed in recent literature for OFDMA multi-hop relay networks. For example, [53] provided dynamic allocation methods for a linear MMR 802.16j network utilizing an effective spectrum efficiency index, which reduced the time-consuming dynamic resource allocation while still maintaining throughput fairness.

[5] implemented a first come first serve (FCFS) scheduling strategy in a two-hop network. The resources were time partitioned and assigned to flows in a manner that uplink and downlink TCP traffic were kept balanced. This adaptive algorithm showed noticeable improvement over static systems.

In [6], centralized sum rate maximization was adopted as a resource allocation method in an OFDMA two hop cellular network. An optimization problem was developed to schedule radio resources with and without power allocation. A mixed integer problem was formulated and simplified by relaxing integer constraints to form a linear program that could be solved via standard methods.

The authors in [7] relied on a graph theoretic approach to jointly optimize routing and subcarrier allocation, maximizing the throughput and assuring that a minimum amount of resources was allocated to each user.

[8] used a heuristic two-hop approach based on a proportional fair algorithm. Another two-hop scheme was proposed in [54] where a TDM frame partitioning was adopted. One TDM partition was devoted for BS to UT/RS transmission followed by a

RS to UT partition. A non-cooperative power allocation game was defined to maximize throughput followed by an iterative solution.

There are numerous solutions available in literature, each with different assumptions, approaches, goals, frame structures and scheduling strategies [5]-[8],[53] and [54]. However, most of them suffer from scalability issues, since they are designed for two-hop deployments only. Although two-hop solutions get most of the benefits from relaying, still relay systems with more hops can find their applications especially under low user load or density scenarios. However, increasing the number of hops leads to increased contention on the resources by the relay operation caused by management and feedback overhead, and traffic relaying. Therefore, improving resource utilization by either reducing the overhead and/or enabling IRRR is necessary for efficient operation of the system.

One of the proposals that considers more than two-hops can be found in [9] where a centralized scheduling framework was developed. The authors formalized an OFDMA optimal centralized scheduling problem and proofed its NP-hardness and approximation-hardness. Heuristic solutions were then presented. Their results indicated a small improvement over [10]. Centralized route selection and scheduling problems were studied in [10]. Maximum SINR, proportional fair, and round robin algorithms were modified to serve in an OFDMA environment, which assigned tones to users based on their end-to-end metrics. Both [9] and [10] assume that the system is able to relay the traffic from slot to slot, i.e., within a single frame, the BS and relays are able to forward the traffic to the designated user. This is not realizable as the relay

receives the control messages first, and only then is able to receive the relayed traffic. Thus, the relay needs at least a single frame to forward the relayed traffic.

In this work, we will develop two realizable scheduling algorithms designed for OFDMA relay enhanced networks. Our algorithms accept one or more relay hop count while reducing problem complexity and overhead requirements. These algorithms were motivated by our result in the previous chapter and in [12], which indicates that applying the same AMC mode on all links towards the destination would not effectively impact the overall performance of the network. With this implementation, we not only attempt to simplify the problem, but also to save in management and feedback overhead.

4.3 Contributions

1. We have developed a novel and simple scheduling algorithm for a multi-hop relay that provides good performance compared to existing scheduling methods.
2. We further developed this algorithm to allow for a simple reuse and power control strategy, which provides a great improvement to the system throughput with a modest increase in complexity.

The solutions introduced in this chapter are traffic-aware. They allocate the resources based on the conditions of the received traffic, unlike in [9] or [10], which assume a constant stream of traffic. Statistical multiplexing is essential to get the benefit from traffic diversity in order to improve the throughput.

Therefore, system throughput cannot be maximized if the system is traffic or queue unaware. [11]

4.4 System description

In a cellular network, it is assumed that each cell is served by a single BS and multiple multi-hop RS's. It is assumed that data are transmitted over two dimensional radio resource units comprising of frames. The scheduler is required to select what link will be activated, whether it is an access link or a relay link, and it should determine what AMC mode will be chosen.

In the previous chapter, we have formulated our problem which needed simplification in order to have an efficient solution. One of the proposed strategies was to adopt a multi-frame concept (see Figure 4.1) which requires the scheduling interval to be long enough to accommodate all hop traffic. This would result in a simplified algorithm, as we will see in the later sections.

Figures 4.1 and 4.2 outline the multi-frame concept, presenting a simple example of a cell with a single BS, one single-hop RS, and one two-hop RS. We have four users: A, B, C and D. A is served directly by the BS. B is served by RS 1, while the remaining, C and D, are served by RS 2. At the first frame in the multi-frame time v, the BS sends the traffic of all users. It sends the control massage for all users, the data traffic for its access user A and the relay traffic for its relayed users B, C, and D. All the relays are in listening mode. In the next frame (τ+1) in the same multi-frame, the BS is not allowed to send any relay traffic, but it can transmit to its access user A.

All RS's will be silent at this frame time. In the next frame, all one-hop RS's are allowed to transmit. They can transmit to their access user B and send the relay traffic of C and D. At the third frame (τ+2), the two-hop RS's are allowed to transmit. In our case RS 2 is allowed to transmit to its users C and D.

Since this is a centralized scheduling scheme, it is expected to have a queue for each flow. We are considering a single traffic class and, hence, we require a single queue for each flow (refer to Figure 4.2). The BS stores the arrival packets of each flow in its queue. While queues will be selected for service based on the proposed scheduling algorithms, the packets in the same queue are emptied on a FIFO basis.

4.5 Weakest Link Multi-Frame Forwarding (WLMF) algorithm

The WLMF assumes a scheduling duration equal to the maximum number of hops. The frames are grouped to form what we refer to as "multi-frame". Referring to the scheduling problem, Equation 3.40, the third constraint can be met when we restrict the transmission activation according to hop count. This means the second hop links are activated in the second frame of a multi-frame, and third hop links are restricted for the third frame in a multi-frame. For the first hop link, the BS can check for available resources and utilize them for its access traffic; however, for relay traffic, the BS is restricted to the first frame in a multi-frame. The key feature of our solution is to *allocate the same resources to all links belonging to the same flow at all frames in a multi-frame*. By doing so, we satisfy the sixth constraint. The first constraint C1-Eqn 3.40 will be guaranteed using a simple policy during the scheduling process

where we define a set denoted by $G(m)$ containing the flows that have packets waiting for transmission. $G(m)$ is updated whenever a queue of a flow changes its state from empty to non-empty or from non-empty to empty, otherwise $G(m)$ remains unchanged.

The second and forth constraints are met if we restrict the transmission to a single node only. With this formulation, satisfying constraint 5 will satisfy constraint 4. Therefore, the problem (Eq. 3.40) has transformed into

$$\max_{X} U(m) = \sum_{\forall i \in G(m)} \sum_{\forall f \in \mathbb{G}^{(f)}} \sum_{\forall t \in \mathbb{G}^{(t)}} x_{i,f,t}(m)\, u_{i,f,t}(m)$$

Subject to (4.1)

$$\sum_{\forall i} x_{i,f,t}(m) \leq 1$$

The problem is a typical GAP which is known to be NP-hard. We will adopt a sub-optimal greedy approach to solve this problem. This greedy approach is optimal in the case of FULL Queue. The solution in this case will be:

$$x_{k,f,t}(m) = \begin{cases} 1 \; if \; k = argmax(u_{i,f,t}(m)) \\ 0 \; otherwise \end{cases} \quad (4.2)$$

where m is the multi-frame index and $u_{i,f,t}(m)$ is the utility of the i^{th} flow

4.6 Utility Functions

Different utility functions can be used. Under the context of multi-hop relay allocation problem, utility functions can reflect the end-to-end performance or link performance. Five scheduling utilities will be studied and presented in Table 1. The first two are end-to-end while the last three are link based. The access link is considered in the last three functions. The utilities under consideration are:

1. End-to-End Maximal throughput (E2EMT)
2. End-to-End Proportional Fairness (E2EPF)
3. Proportional Fairness (PF)
4. Maximum Weakest Link Throughput (MT)
5. Maximum Fairness (MF)

Optimization Criterion	**Utility** $u_{i,j,s}(m)$	**Goal**
End-to-end maximal Throughput	$\gamma_{h_i} c_{i,j,s}(m)$	The flows are selected based on the end-to-end spectral efficiency
End-to-end Proportional fairness	$\gamma_{h_i} \frac{c_{i,j,s}(v)}{\bar{r}_i(v)}$	The flows are selected such that each flow costs the same radio resources
Weakest link maximal throughput	$c_{i,j,s}(m)$	The selection is based on the weakest link channel quality
Weakest link Proportional fairness	$\frac{c_{i,j,s}(v)}{\bar{r}_i(v)}$	The flow is selected such that each flow costs the same access radio resources
Maximal fairness	$\frac{1}{\bar{r}_i(v)}$	The flows are selected such that each flow get the same throughput

Table 4.1 Different utility functions for WLMF and WLDF algorithms

In the case of multi-hop communication and with no IRRR, the weakest link has the biggest impact on the end-to-end spectral efficiency, i.e., the link with lowest SINR, has always the biggest impact on the spectral efficiency of multi-hop link since it is always less than the spectral efficiency of the weakest link. The spectral efficiency of a slot $s = (f, t)$ over a link j carrying data targeting a flow k can be written as $c_{k,j,f,t} = (1 - BLER_{k,j,f,t})C\_MAX_{k,j,f,t}$, where $C\_MAX_{k,j,f,t}$ is the maximum spectral efficiency of slot (f, t) based on the estimated SINR and determined by AMC mode selection and $BLER_{k,j,f,t}$ is the block error rate which is a function of SINR and the AMC mode too.

In our model, the spectral efficiency of the weakest link, $\min_{j\in\Im(k)}(C\_MAX_{k,j,f,t})$ will determine the spectral efficiency of all links $\{j: j \in \Im(k)\}$ along the path from the BS to the destination flow k. The maximum spectral efficiency of each link j along the path from BS to k equals $\min_{j\in\Im(k)}(C\_MAX_{k,j,f,t})$.

The end-to-end spectral efficiency, $c_{k,f,t}$ can be calculated by computing the end-to-end $BLER_{k,f,t}$ and applying it to the same spectral efficiency formula which can be written as:

$$c_{k,f,t} = \gamma_k(1 - BLER_{k,f,t})C\_MAX_{k,f,t} \tag{4.3}$$

where γ_k is the utilization of resources for flow k. Under the assumption of independent link qualities, $BLER_{k,f,t}$ can be calculated from the individual $BLER_{k,j,f,t}$

as $BLER_{k,f,t} = 1 - \prod_{\forall j \in \Im(k)}(1 - BLER_{k,j,f,t})$. Note that $BLER_{k,j,f,t}$ is based on $C\_MAX_{k,f,t}$ determined by the AMC mode of the weakest link:

$$C\_MAX_{k,f,t} = \min_{j \in \Im(k)}(C\_MAX_{k,j,f,t}) \tag{4.4}$$

$BLER_{k,f,t}$ is upper bounded by $h_{max} * \max_{j \in \Im(k)}(BLER_{k,j,f,t})$ which is the case when all the link of the same worst qualities: $BLER_{k,f,t} < h_{max} \max_{j \in \Im(k)}(BLER_{k,j,f,t})$. Accordingly, the spectral efficiency can be bounded as:

$$c_{k,s} > \left(1 - h_{max} * \max_{j}(BLER_{k,j,s})\right) \gamma_k \min_{j}(c_{k,j,s}) \tag{4.5}$$

Thus $c_{k,s}$ can be approximated by:

$$c_{k,s} = a\gamma_k \min_{j} c_{k,j,s}\ ,\ \left(1 - h_{max} * \max_{j}(BLER_{k,j,s})\right) < a < 1 \tag{4.6}$$

If the BLER is very small compared to 1, we can approximate a to be equal to 1. γ_k is a variable represents the utilization associated with flow from BS to UT k which represents how many times a slot is effectively used for transferring data. For example, in the case of three hops and three frames multi-frame, the single hop flows will be able to use the resources three times while the 2 and 3-hop flows will be able to utilize the resources only once. The factor γ_k is used to reflect resource utilization as follows

$$\gamma_k = \begin{cases} 1 & \text{if } h_k = 1 \\ \dfrac{1}{h_{max}} & \text{if } h_k \neq 1 \end{cases} \tag{4.7}$$

where h_{max} is the maximum number of hops. Here, we can notice the utilization becomes very low as h_{max} increases. To improve the system utilization, we let the BS use the unutilized resources scheduled for the relayed users during the third frame and a higher numbered frame in a multi-frame. This requires another algorithm run on the third frame and for the access traffic at the BS only. In this case, the utilization factor will be:

$$\gamma_k = \frac{1}{h_k} \tag{4.8}$$

Throughput based utilities over a link can be expressed as a function of the spectral efficiency of the link $c_{i,j,f,t}(m)$. Link based throughput utility is the weakest link utility, which equals $\min_{j\in\Im(i)}[c_{i,j,f,t}(m)]$.

In general, we can define the weakest link maximum throughput utility as $u_{i,f,t}(m) = \min_{j\in\Im(i)}[u_{i,j,f,t}(m)]$, where the utility $u_{i,j,f,t}(m)$ is equal to

$$u_{i,j,f,t}(m) = \begin{cases} \gamma_k c_{i,j,f,t}(m) & \text{for end. to. end throughput utility} \\ c_{i,j,f,t}(m) & \text{for link based throughput utility} \end{cases} \tag{4.9}$$

The same can be said about proportional fairness utilities; see Table 1 for details.

4.7 Weakest Link Duo-Frame Forwarding (WLDF) algorithm

This algorithm enhances the utilization of relay links by reusing the same radio resources on every second hop simultaneously. In this case, two frames (even and odd) are needed. The utilization is improved since all links are simultaneously utilized by every second hop link. The utilization factor γ_k improves in this case:

$$\gamma_k = \begin{cases} 1 & \text{if } h_k = 1 \\ 1/2 & \text{otherwise} \end{cases} \tag{4.10}$$

Regardless of the number of hops, the minimum utilization factor will be 1/2. Here we do not restrict ourselves to any number of hops. Thus, this algorithm is scalable. The same utility function presented in the previous section can be used along with the same solution method and format.

4.8 Improved Solution

After finding the solutions $\{x_{k,f,t}(m)\}$ using Equation 4.2 for all flows, the scheduler determines the AMC mode for each flow, which is dependent on link quality. The AMC for each flow will be based on the weakest link and will be the same for all hops. Therefore, the solution can be written as

$$\mathrm{r}_{k,f,t}(m) = \begin{cases} \min\limits_{j \in \Im(k)} [c_{k,j,f,t}(m)] & \text{if } k = \underset{\forall i \in G(m)}{\operatorname{argmax}} \left[\min\limits_{j \in \Im(i)} [u_{i,j,f,t}(m)] \right] \\ 0 & \text{Otherwise} \end{cases} \tag{4.11}$$

where $\mathrm{r}_{k,f,t}(m)$ is the instantaneous spectral efficiency of flow k on slot (f,t). In scattered slots assignment, which introduces a significant overhead. Thus, a reasonable

The resources, as we stated earlier are two dimensional, and comprise a number of sub-carriers. The method by which the subcarriers are assigned to a slot is called permutation. In general, there are two types of permutations: distributed and contiguous.

The contiguous permutation is referred to as Adjacent Subcarrier Permutation (ASP), which combines a number of adjacent subcarriers to form sub-bands. This improves multi-user diversity gain [1].

The second type is referred to as Distributed Subcarrier Permutation (DSP). Here, the sub-carriers are not adjacent. Instead, they are picked in a distributed fashion to provide some frequency diversity gain, ensuring a more robust permutation type. With ASP, the utility on some or all sub-channels (or sub-bands) is evaluated whereas only one value is required with DSP. The difference between the two cases is that in ASP, every sub-channel is scanned and the algorithm is applied on each. In both cases, when a slot is allocated to a flow, the adjacent slot allocation continues until either no further packet is available, or until there are no more slots to be allocated.

4.9 Power Control

Power control is needed in WLDF to allow simultaneous transmission between relays. The interference between simultaneously active links can be high. Therefore, in order to reduce the interference, we want to find the optimal transmission power on every link towards the destination such that the throughput is maximized. It can be shown that the optimal power allocation is the one that results in equal spectral efficiency on

all links. Similarly, the received SINR values are equal for each link. For the i^{th} path, the received SINR of link l is given by

$$SINR_l = \frac{P_{r_l}}{N + I_l} \tag{4.12}$$

where P_{r_l} is the received power, N is the noise power and I_l is the received interference. Two types of interference are experienced by link l: one is associated with the links belonging to the same flow while the other is from the surrounding cells; that is $I_l = I_l^i + I_l^o$, where I_l^i is the interference experienced by link l from the links on the path of the i^{th} flow and I_l^o is the interference from other sources. The interference I_l^i has to be estimated in order to optimize power allocation. We can write I_l^i as

$$I_l^i = \sum_{\forall l' \in G_l^i(l)} P_{t_{l'}} \mathbb{h}_{l'l} \tag{4.13}$$

where $G_l^i(l)$ is the set of links in the path of flow i that interferes with link l, $P_{t_{l'}}$ is the transmitted power of link l' and $\mathbb{h}_{l'l}$ is the channel gain from transmitter of link l' to the receiver of link l. In this case, $SINR_l$ can be written as

$$SINR_l = \frac{P_{t_l} \mathbb{h}_{ll}}{N + I_l^o + \sum_{\forall l' \in G_l^i(l)} P_{t_{l'}} \mathbb{h}_{l'l}} \tag{4.14}$$

where P_{t_l} is the transmitted power by the transmitter of link l and $\mathbb{h}_{ll}$ is the channel gain from transmitter to the receiver of link l. If flow i is of h_i hops, we solve h_i equations (one equation for each link) to find the value of transmitted power $\{P_{t_l}\}$ of each link. We then rewrite the above equation as

$$1 = \frac{P_{t_l}\mathbb{h}_{ll}^{\#}}{N_l + I_l^o + \sum_{\forall l' \in G_I^I(l)} P_{t_{l'}}\mathbb{h}_{l'l}} \tag{4.15}$$

where $\mathbb{h}_{ll}^{\#} = \frac{\mathbb{h}_{ll}}{SINR_l}$. Since all links have the same SINR, we set $\mathbb{h}_{ll}^{\#} = \frac{\mathbb{h}_{ll}}{SINR}$ for all links.

Allowing $W_l = N_l + I_l^o$, the equation can be rewritten as

$$W_l + \sum_{\forall l' \in G_I^I(l)} P_{t_{l'}}\mathbb{h}_{l'l} - P_{t_l}\mathbb{h}_{ll}^{\#} = 0 \tag{4.16}$$

This will produce a system of linear equations which can be solved to get the values $\{P_{t_l}\}$ as function of $\mathbb{h}_{l'l}$ and $\mathbb{h}_{ll}^{\#}$ and consequently as function of SINR. For the three hop case, we solve the system of linear equations to get

$$\begin{aligned}
P_{t_1} &= \frac{W_1\mathbb{h}_{33}^{\#} + W_3\mathbb{h}_{31}}{\mathbb{h}_{11}^{\#}\mathbb{h}_{33}^{\#} - \mathbb{h}_{13}\mathbb{h}_{31}} = \frac{W_1\frac{\mathbb{h}_{33}}{SINR} + W_3\mathbb{h}_{31}}{\frac{\mathbb{h}_{11}\mathbb{h}_{33}}{SINR^2} - \mathbb{h}_{13}\mathbb{h}_{31}} \\
P_{t_2} &= \frac{W_2}{\mathbb{h}_{22}^{\#}} = \frac{W_2}{\frac{\mathbb{h}_{22}}{SINR}} \\
P_{t_3} &= \frac{W_3\mathbb{h}_{11}^{\#} + W_1\mathbb{h}_{13}}{\mathbb{h}_{11}^{\#}\mathbb{h}_{33}^{\#} - \mathbb{h}_{13}\mathbb{h}_{31}} = \frac{W_3\frac{\mathbb{h}_{11}}{SINR} + W_1\mathbb{h}_{13}}{\frac{\mathbb{h}_{11}\mathbb{h}_{33}}{SINR^2} - \mathbb{h}_{13}\mathbb{h}_{31}}
\end{aligned} \tag{4.17}$$

We can increase the SINR until either $\frac{\mathbb{h}_{11}\mathbb{h}_{33}}{SINR^2} \geq \mathbb{h}_{13}\mathbb{h}_{31}$ or until either one of the values of $\{P_{t_l}\}$ reaches maximum.

$$\text{SINR} = \min\left(\sqrt{\frac{\mathbb{h}_{11}\mathbb{h}_{33}}{\mathbb{h}_{13}\mathbb{h}_{31}}}, SINR\left(P_{t_{l_{max}}}\right)\right) \tag{4.18}$$

After we have identified the highest SINR, we plug it back to Equations (4.17) to get the values $\{P_{t_l}\}$.

Note that this procedure creates a constraint on the allocation power of a future assignment of the same slot should be considered in the allocation problem. Therefore, the values $\{c_{i,j,f,t}(m)\}$ would not be based on equal power assumption; rather, they will be based on the power allocation procedure mentioned earlier.

4.10 Performance Simulation

4.10.1 Simulation Models and Parameters

Computer simulation is our method of choice to evaluate the performance of our system. System assumptions and parameters are based on the recommendations of IEEE 802.16-j Task Group [14]. Simulations will be carried out using MATLAB to study throughput and fairness performance of our algorithms. MATLAB is a powerful simulation tool especially in modeling the physical layer and analyzing the results. We will use the parameters presented in Table 3-1. The simulation of each deployment is repeated at least 20 times, 50 drops for each time and with 200 frames per drop. The number of simulation runs met the recommendations of IEEE802.16 task groups. When the variance of the result is large, we will increase the number of experiments to reduce the variance; and hence we will reduce the confidence interval to enhance the accuracy of the results.

The large scale channel impairments are kept fixed during the simulation. However, small scale fading will vary from frame to frame. Poisson arrival is assumed at different demand levels. Usually Poisson is used to model voice call or messages arrival [86].

We simulate the performance of a cell surrounded by 18 interfering cells. Full queue is assumed at each transmitting node in the surrounding cells. Thus, the nodes inside the center cell experience a continuous interference from the interfering nodes in the surrounding cells. The cell is partitioned into three sectors, the available resources are distributed equally on each sector in a manner that no overlap between the resources is allocated to any sector. There are two examples of relay deployments: two-hop deployment containing six one-hop RS's (Case 2, Figure 3.1) and a three hop deployment with a total of 18 RS's, 6 one-hop RS's and 12 two-hop RS's (Case 5, Figure 3.1), in addition to the legacy (single-hop) system. A frequency reuse of one in each cell is assumed; i.e. each one of the surrounding cells uses the same radio resources. The same deployment is assumed for all cells.

As in section 3.2.6, we assume a NLOS link between the BS or an RS, their subordinate UT's, and a LOS link between the BS or a RS, and another RS if they have a direct father-child relationship. Otherwise, a NLOS link is assumed. In the LOS cases, we will use the modified IEEE 802.16 type-D path loss model defined in [14], which assumes a suburban flat terrain environment. The standard deviation of lognormal shadowing is 3.4dB. In the case of NLOS, whether it is between BS/RS and UT or between BS/RS and a subordinate RS, the modified IEEE 802.16 path loss model type B assumes intermediate suburban path loss conditions [14], i.e. moderate tree densities with neither flat nor hilly terrain. The lognormal shadowing of 9.6 dB standard deviation is assumed.

To model the multipath fading channel, a tapped delay line is used with different parameters based on measurements for each link type. SUI 1 is chosen for the links between RS's and between the BS and the RS's. For the links to UT's, we will consider ITU Ped-A channel model. Refer to Table 3.1 for the other model assumptions.

4.10.2 Results and discussions:

[9] and [10] presented their solutions for cellular networks with two or more relay hops. [9] showed the superiority of their solution compared to [10]. Therefore, we will compare our algorithms to the one named ArgMAX proposed in [9] since the other solutions are not realizable as it assumes the relay traffic passes to the destination within a single frame. Although ArgMAX algorithm is also not realizable, it can be modified to be done in multiple frames. We will consider the reference algorithm under the two relay deployments mentioned in the previous section. The two reference cases are referred to as CPF-2H and CPF-3H, where CPF refers to Centralized Proportional Fair algorithm. Basically, the algorithm assigns the resources according to the number of users in each hop level and forwards the traffic based on proportional fair flow utility, such that the flow with the highest utility will be granted the opportunity. The flow will be assigned if there are enough resources on all links towards the path to the destination. Based on link quality, optimal AMC mode will be selected and the required number of resources will be calculated.

Figure 4.3 and 4.4 present a comparison between WLMF, WLDF and the reference 2-hop and 3-hop CPF system in terms of average cell throughput and fairness. We also present the results for single hop (legacy) system for comparison. Figure 4.3 shows some improvement for relay systems compared to a single hop system at low network load. The number of users, as well as their demand, contributes to network load; increasing either one increases network load. As a general observation, at low load conditions, increasing hop count improves network throughput due to improved users' connectivity, i.e., improved outage performance, whereas increasing the load via increasing the demand or number of users, reduces this improvement to the point that increasing hop count becomes a negative effect on system throughput. Figure 4.3 illustrates that at high demand, the two-hop system performs better than the three hop cases except for a WLDF system, which performs the best at high demand due to improved utilization compared to the WLMF or reference system, and exhibits improved connectivity compared to the two hop cases. Hop count has a direct impact on throughput fairness performance as increasing hop count increases connectivity, which means reduced outage and this reflects directly on the fairness performance under the usage of proportional fairness objective function, which tends to give each user equal opportunity (resources).

Although simple and easy to implement, our system did not experience a noticeable performance reduction compared to the reference system. In fact, in relatively low to medium load, our system shows some throughput improvement. This improvement came from the flexibility in assigning the resources, unlike the reference

system, which statically divided the resources according to the number of users without considering the arrival traffic conditions. In our case, the system assigns the resources according to the scheduled traffic. Increasing the number of users or their demand reduces the impact of static assignment. In this case, the reference system has a small throughput gain compared to ours due to better utilization of system resources by optimizing AMC selection.

Due to its reuse capability, WLDF was able to achieve around 27% throughput improvement compared to WLMF, at high load conditions. This improvement comes at a price of a small reduction in fairness (around 3.5%). In fact, throughput fairness improvement is highly affected by the outage performance of the system, in addition to the scheduling algorithm. As illustrated in Figure 4.4, by increasing hop count, we achieved better throughput fairness. There is no significant difference in fairness between our WLMF algorithm and the reference system. WLDF, however, experienced a slight fairness reduction at high load conditions compared to the WLMF-3H case. The reason for this reduction is possibly the increase in outage probability due to high interference conditions.

The CDF of an average user throughput is plotted in Figure 4.5. This figure is very interesting as it describes what is happening in the system. At low load, see Figure 4.5-a, the 10th percentile rate of 0.235Mb/s, a number very close to the average demand has been achieved by all relay scenarios. While the single hop case suffers from 0.15 outage probability, the outage reduced in the two-hop case to 0.05. At a higher load, Figure 4.5-b depicts the benefit of increasing the relay hop count, as the

10^{th} percentile rate of 0.95Mb/s has been achieved by the three hop cases. This rate is reduced to 0.5 for the two hops cases and reduced further to 0 for the single hop system. With reference to Figure 4.5-d, the 10^{th} percentile rate of 3-hop flows at the higher network load reduced to about 0.2Mb/s, while 2-hop cases did not exceed 0.1Mb/s. The single-hop system in this case still struggles at 18% outage.

Figures 4.5-b and 4.5-c show the same performance curves with different rate values. Multiplying demand by number of users gives the same average of network total load. As the load becomes higher, interesting behavior is noticed at all curves; this behavior is observed as tied to hop count. With a rate below 0.3Mb/s, the three hop cases were able to support more users whereas systems with lower hop count could not support more users, but they were able to support higher rates. The two-hop cases supported more than 0.4Mb/s. Higher rates are supported by the single-hop case, which was able to deliver more than 0.78Mb/s, but with much less number of users. The WLDF algorithm has the behavior of WLMF-3H at low rates with little performance loss. It improved on the performance of WLMF-2H at higher rates, and was able to support more than 0.45Mb/s to about 50% of the users.

In Figure 4.6, the average user throughput for each user has been shown as scatter points. The curves in Figure 4.6 are the 3^{rd} polynomial degree fitting of the average users' throughput vs. the distance from the BS. The advantage from relaying in improving the connectivity of the system is clear where we see the curve close to the edge of the cell being comparatively higher than that of the legacy system.

However, the relay system suffers at high load as it cannot cope with the increased traffic due to the contention on resources by the relaying operation.

In all the plots, the tails of the curves of 3-hop cases were higher, indicating better edge performance. Figures 4.6-a and 4.6-c show the scatter points concentrated around a line that is equal to the average user demand meaning that the schedulers were able to support that demand except for the single hop case where a large portion of edge users could not be supported due to outage probability; it is shown as a decline of the fitting curve close to the edge of the cell. Increasing the demand increases the scattering indicating that the network is reaching its capacity limits.

The large variance in the throughput value is due to difference in the received SINR as a result of the distance from the relay and closeness to the edge of sector boundaries. As the node becomes far from the relay and closer to the boundary we expect a reduction in the received signal power and an increase in interference power which reduces the SINR tremendously causing a large variation in throughput performance of users even if they are at the same distance from the BS.

Here we are considering the average of the results to estimate their central tendency. However, when a specific application with certain requirements is considered, we may carry out the analysis with a minimum acceptable QoS requirement. The utility function can be designed to assure that a minimum acceptable QoS requirement is satisfied. This can be a valuable extension to our work.

Impact of scheduling algorithms is depicted in Figures 4.7 and 4.8. Figure 4.7 shows the throughput performance while Figure 4.8 shows the throughput fairness

performance. The performance of maximum throughput fair (MF) scheduler is depicted in Figures 4.7-a and 4.8-a, showing the improvement in both throughput performance and fairness with increasing relays hop count. Improvement in throughput performance was noticeable with adopting PF scheduler (Figures 4.7-b and 4.8-b) in all cases except for the WLMF-3H case, which did not experience a significant improvement. This led to a reduced performance compared to WLMF-2H at high load. E2EPF algorithm (Figures 4.7-e and 4.8-e) provided small improvement over PF algorithm with a significant reduction in fairness.

Maximum throughput (MT) scheduler tries to maximize the throughput by scheduling the resources to the flow that has the best quality. In legacy systems, link quality maps directly to flow quality, unlike relay systems where we have to distinguish between link quality and effective end-to-end flow quality. We used spectral efficiency (SE) as our link throughput measure and End-to-End Spectral Efficiency (ESE) as our end-to-end flow throughput measure.

Schedulers based on maximum lowest link SE are referred to as Maximal Throughput (MT) while those based on ESE are referred to as Maximal End-to-End Throughput (E2EMT). Both of these scheduling strategies achieved the highest possible throughput as illustrated in Figure 4.7-c, d. With reference to figure 4.7-c, by implementing the link maximum throughput strategy, a large performance gain at high load was experienced by the legacy system at the expense of reduced throughput fairness. In low load conditions, this was not the case; in fact, the relay system was able to get some improvement due to its ability to support more users at low rates. In

contrast, end-to-end maximal throughput implementation improved relay systems to be equal or better than the legacy system with much better fairness. This was true at realizable traffic load, but it was not the case under infinite queue assumption, where a considerable fairness gain was noticed. The reasoning behind this behavior is that with end-to-end maximal throughput, at infinite queue, the E2EMT scheduler would not forward any more relayed traffic because the relayed flow would always have a reduced ESE compared to the SE achieved by the one hop flows. Since the number of single hop users in relay cases is much less than those in single-hop cases, it is more likely to find more high SE single flows in a single-hop cell compared to the relay scenarios.

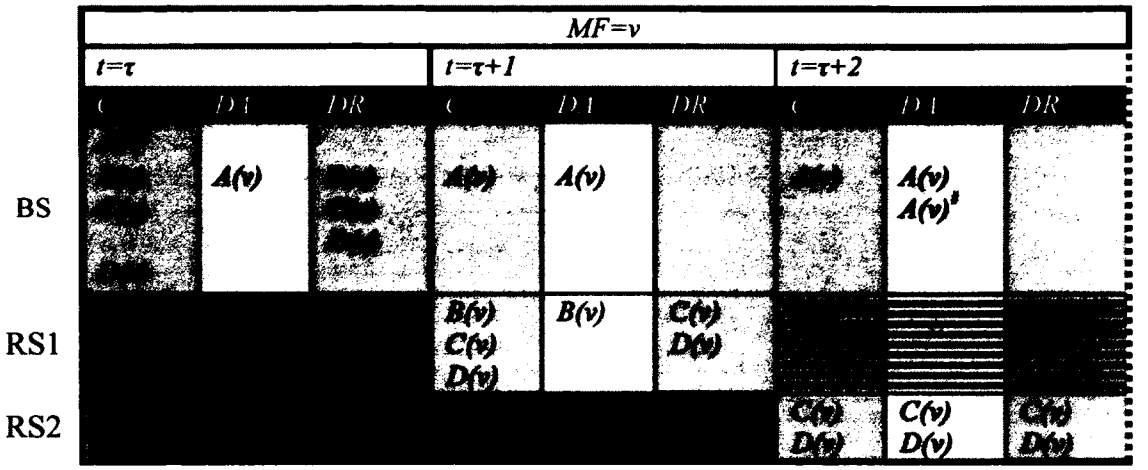
As we are assuming pedestrian users, the system can benefit from ASP. Figure 4.9 shows the effect of selecting the type of permutation on the throughput. At high load, around 11.6% improvement was achieved by ASP in a single-hop system. The improvement reduced to 8.8% in WLMF-2H, 4.1% in WLMF-3H and 7.6% in WLDF. The reason behind this reduction was partially due to the feedback delay in reporting the best band. In addition, for the WLMF-3H, the boost in SINR by ASP did not result in a significant throughput gain, since most users reached the SE limit. This was the same reason behind the insignificant difference between an MF scheduler and a PF scheduler for WLMF-3H. Due to its reduced SINR performance caused by the increased interference, WLDF was able to benefit more from ASP. Reducing the load would reduce the impact of the permutation type to the point that there would be no improvement at all.

4.11 Conclusions

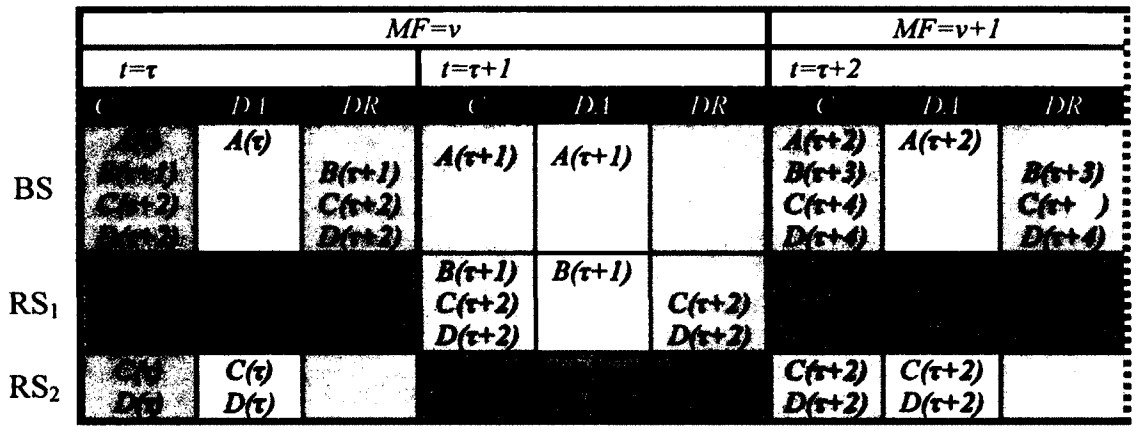
We have proposed a simple and efficient solution for multi-hop relay scheduling, referred to as WLMF. It is extremely simple to implement, requires very limited resources, and was observed to provide comparable performance to the ArgMAX algorithm reported in [9].

With a simple power control mechanism, we have improved on WLMF and proposed WLDF that provided a better utilization due to its ability to reuse the resources. This system shows a good improvement in throughput performance. We noticed that increasing relay hops improved the fairness regardless of which scheduling utility was used, and this is linked directly to outage performance.

With lower network demand, we noticed improved performance, unlike the high load cases which show that WLMF systems suffer from the contention on system radio resources between relay links and access links leading to a reduction in spectral efficiency. WLDF improved this situation by partial reuse of radio resources. This would suggest that WLMF and comparable systems are a good choice to enhance the coverage of the system when the number of users is low, and/or when users demand is low. To achieve higher throughput, we will improve on the IRRR capability further as proposed in our next chapters.



(a) WLMF



(b) WLDF

C.1	Control transmission of Access traffic
D.1	Data transmission of Acces traffic
DR	Data transmission of Relay raffic
$X(v)$	Data or Control belongs to UT_X to be received during Multi-frame time v, where X=A, B, C or D
$X(\tau)$	Data or Control belongs to UT_X to be received during frame time τ, X=A, B, C or D
	Listening Mode
	Wasted Resources

Guide

The BS can utilized unused resources scheduled for flow C

Figure 4.1 Multi-Frame concept for (a) WLMF and (b) WLDF algorithms

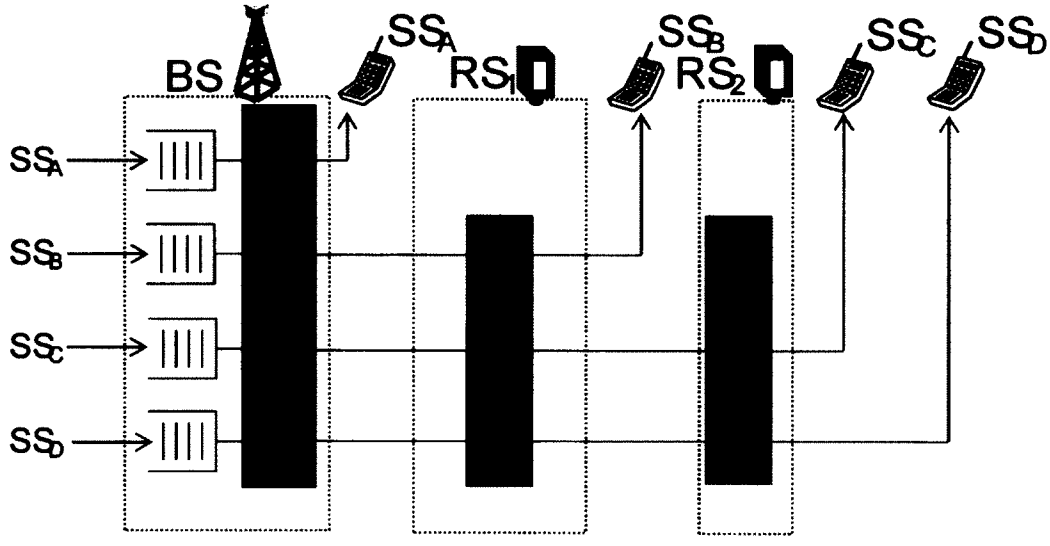


Figure 4.2 Queue Diagram of WLMF and WLDF system

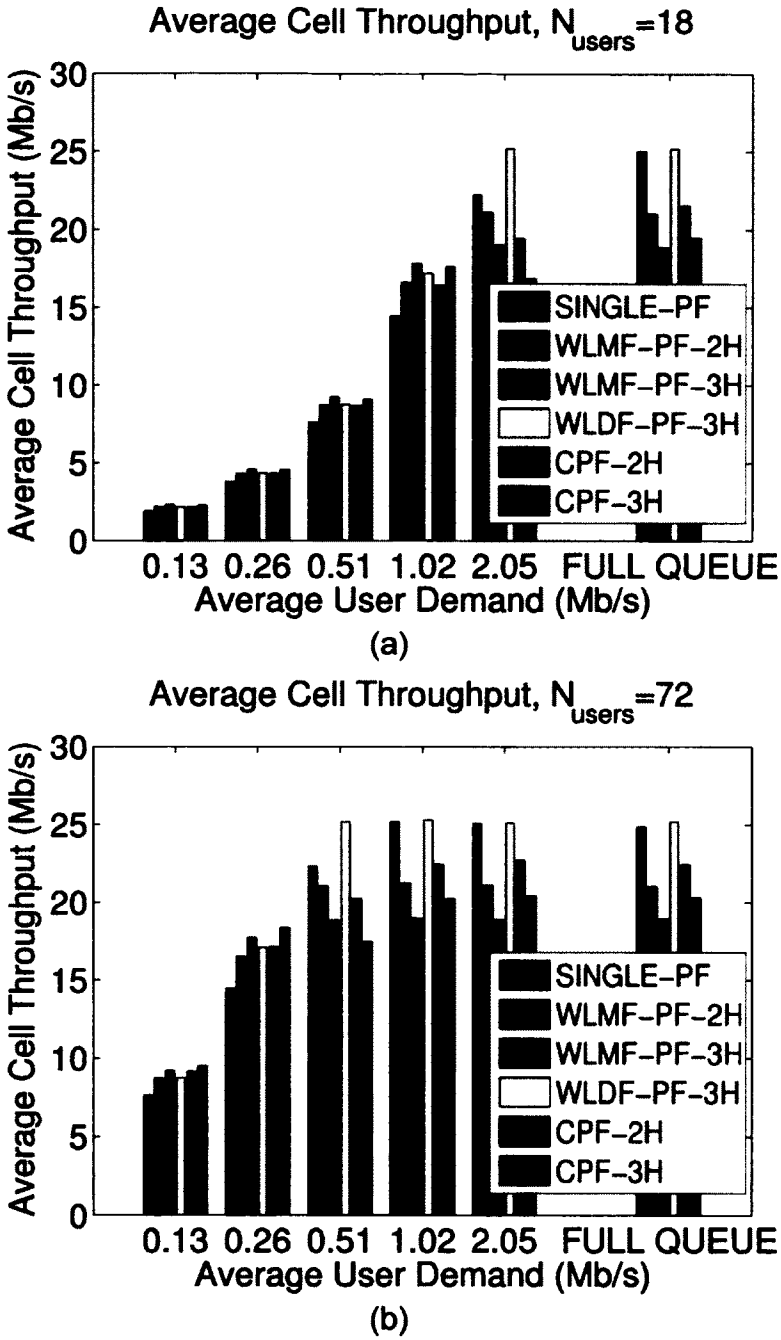


Figure 4.3 Average Cell Throughput vs. users demand with different number of users per cell: (a) number of users equal to 18, (b) number of users equal to 72.

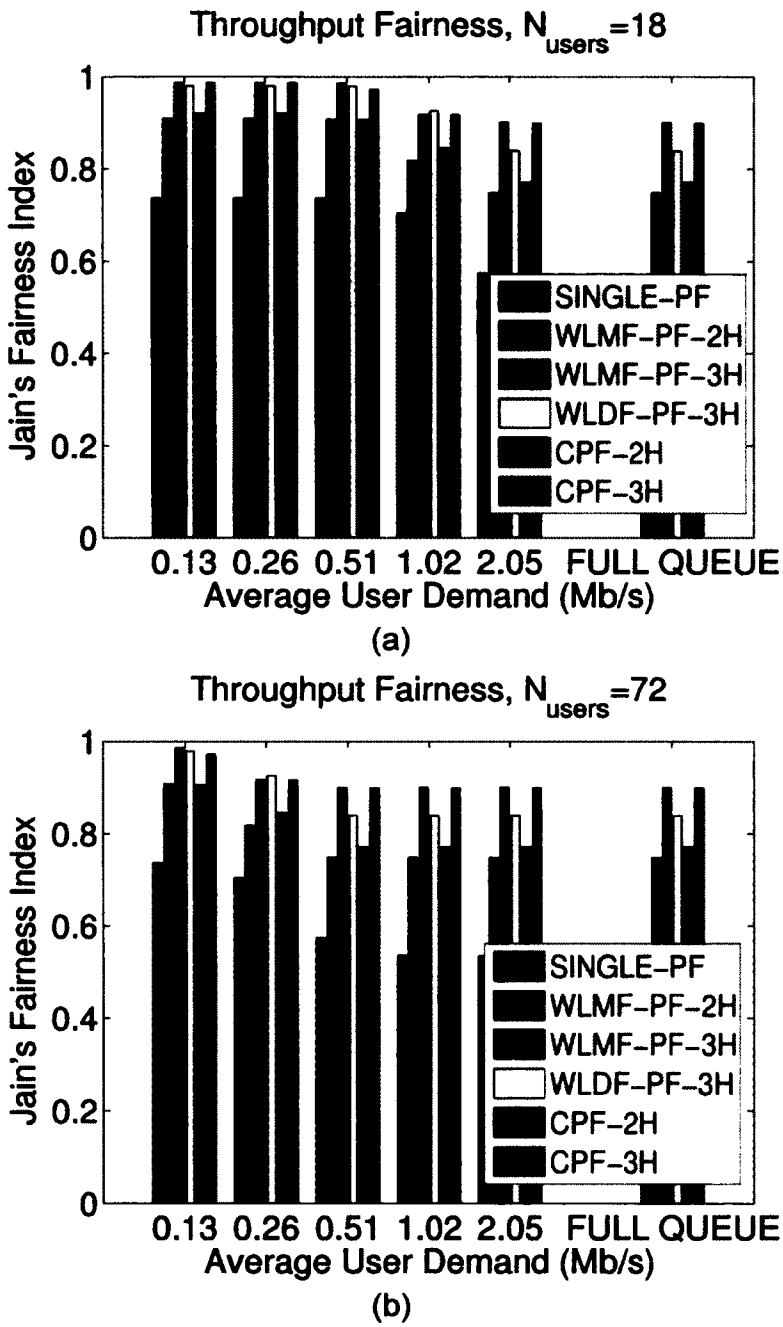
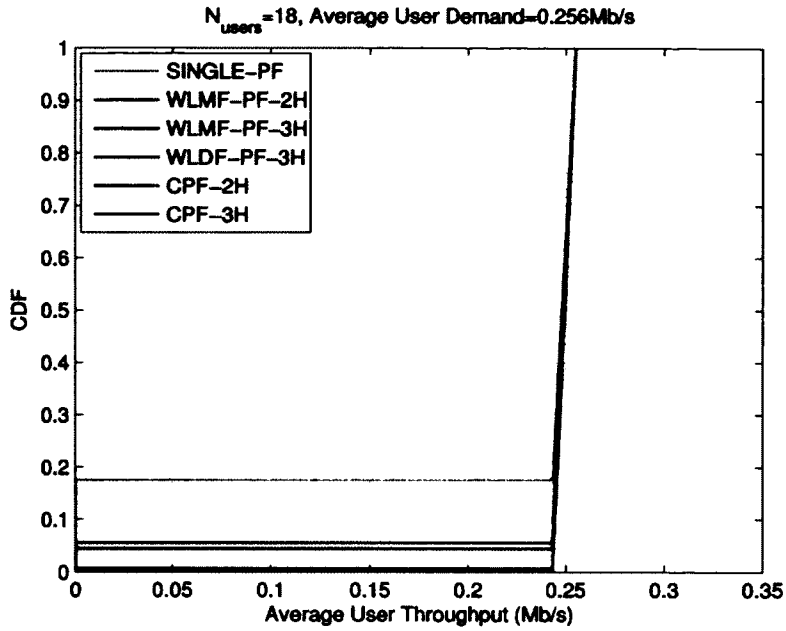
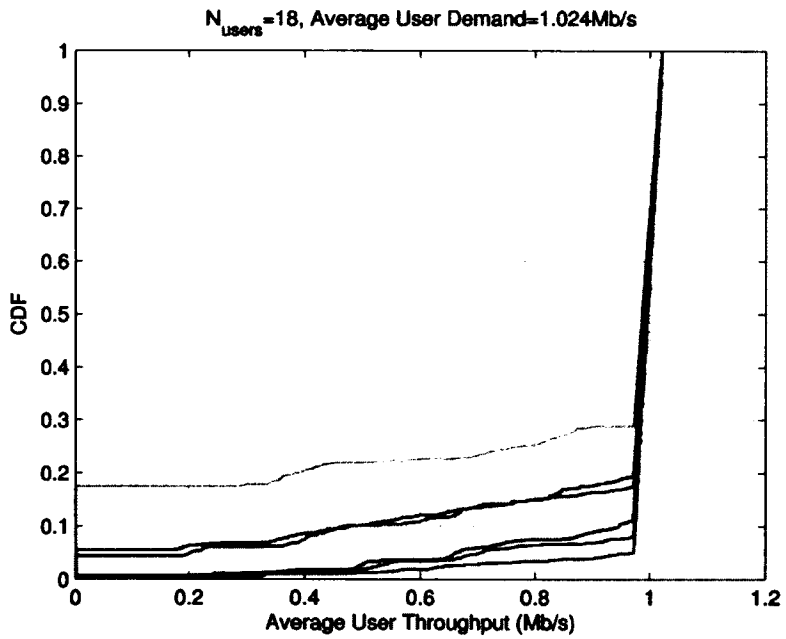


Figure 4.4 Average Throughput Fairness vs. users demand with different number of users per cell: (a) number of users equal to 18, (b) number of users equal to 72.

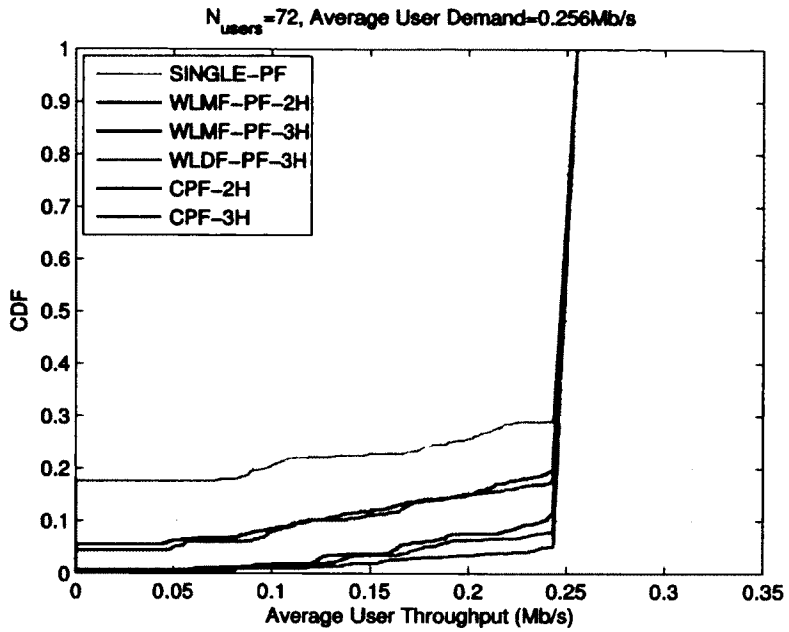


(a)

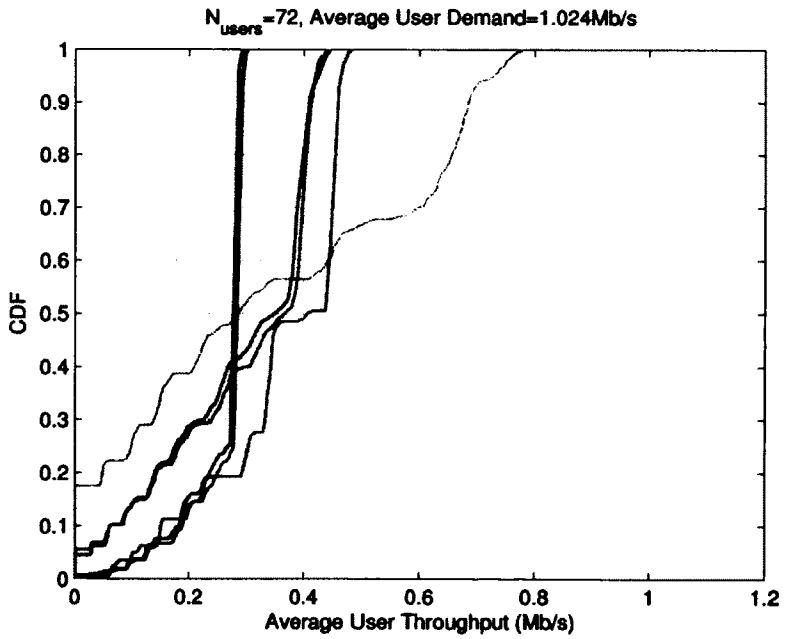


(b)

Figure 4.5 CDF of Average Users throughput: (a) N_{users}= 18, average demand equals 0.256Mb/s (b) N_{users}= 18, average demand equals 1.024Mb/s, (c) N_{users}= 72, average demand equals 0.256Mb/s (d) N_{users}= 72, average demand equals 1.024Mb/s.

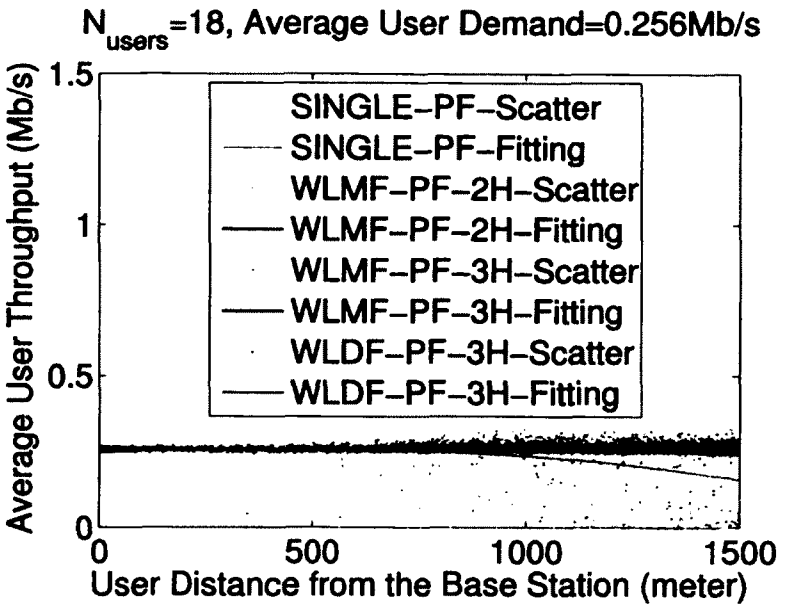


(c)

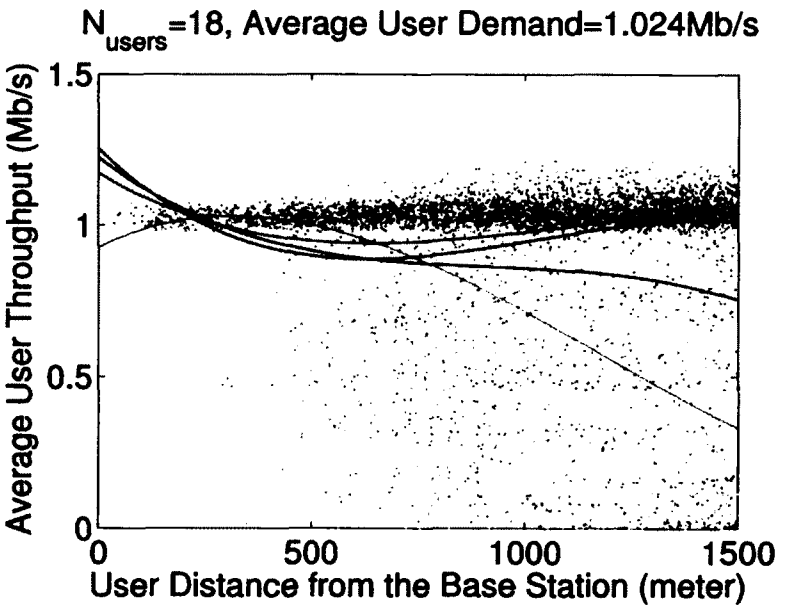


(d)

Figure 4-5 (Cont.) CDF of Average Users throughput: (a) N_{users}= 18, average demand equals 0.256Mb/s (b) N_{users}= 18, average demand equals 1.024Mb/s, (c) N_{users}= 72, average demand equals 0.256Mb/s (d) N_{users}= 72, average demand equals 1.024Mb/s.

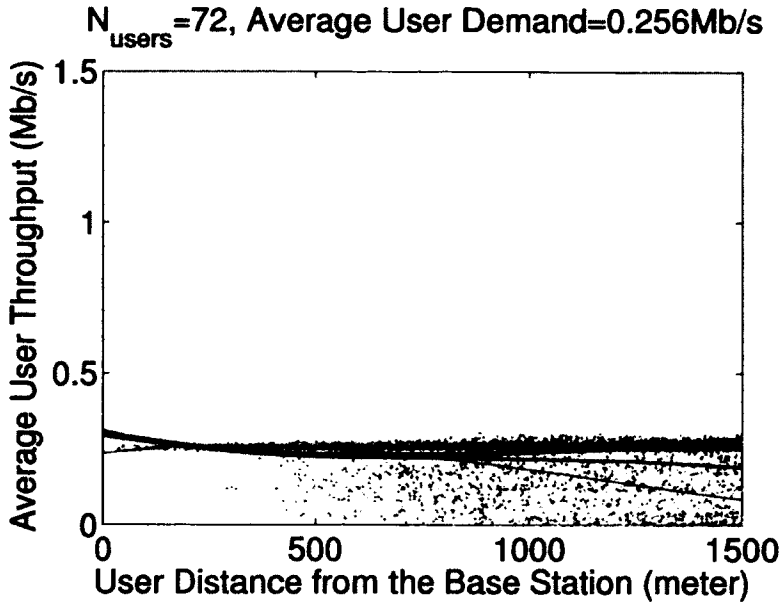


(a)

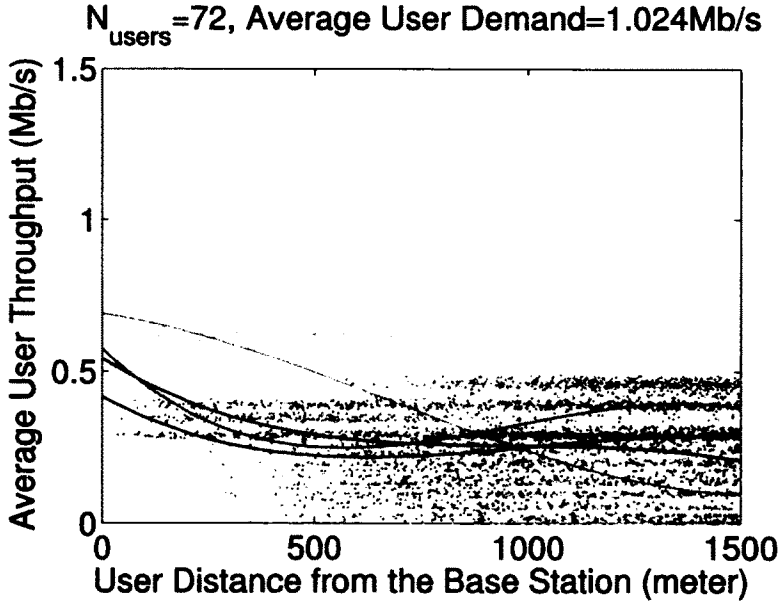


(b)

Figure 4-6 Scatter plot and fitting curves for Average Users throughput vs. distance: (a) N_{users}= 18, average demand equals 0.256Mb/s (b) N_{users}= 18, average demand equals 1.024Mb/s, , (c) N_{users}= 72, average demand equals 0.256Mb/s (d) N_{users}= 72, average demand equals 1.024Mb/s.



(c)



(d)

Figure 4-6 (Cont.) Scatter plot and fitting curves for Average Users throughput vs. distance: (a) N_{users}= 18, average demand equals 0.256Mb/s (b) N_{users}= 18, average demand equals 1.024Mb/s, , (c) N_{users}= 72, average demand equals 0.256Mb/s (d) N_{users}= 72, average demand equals 1.024Mb/s.

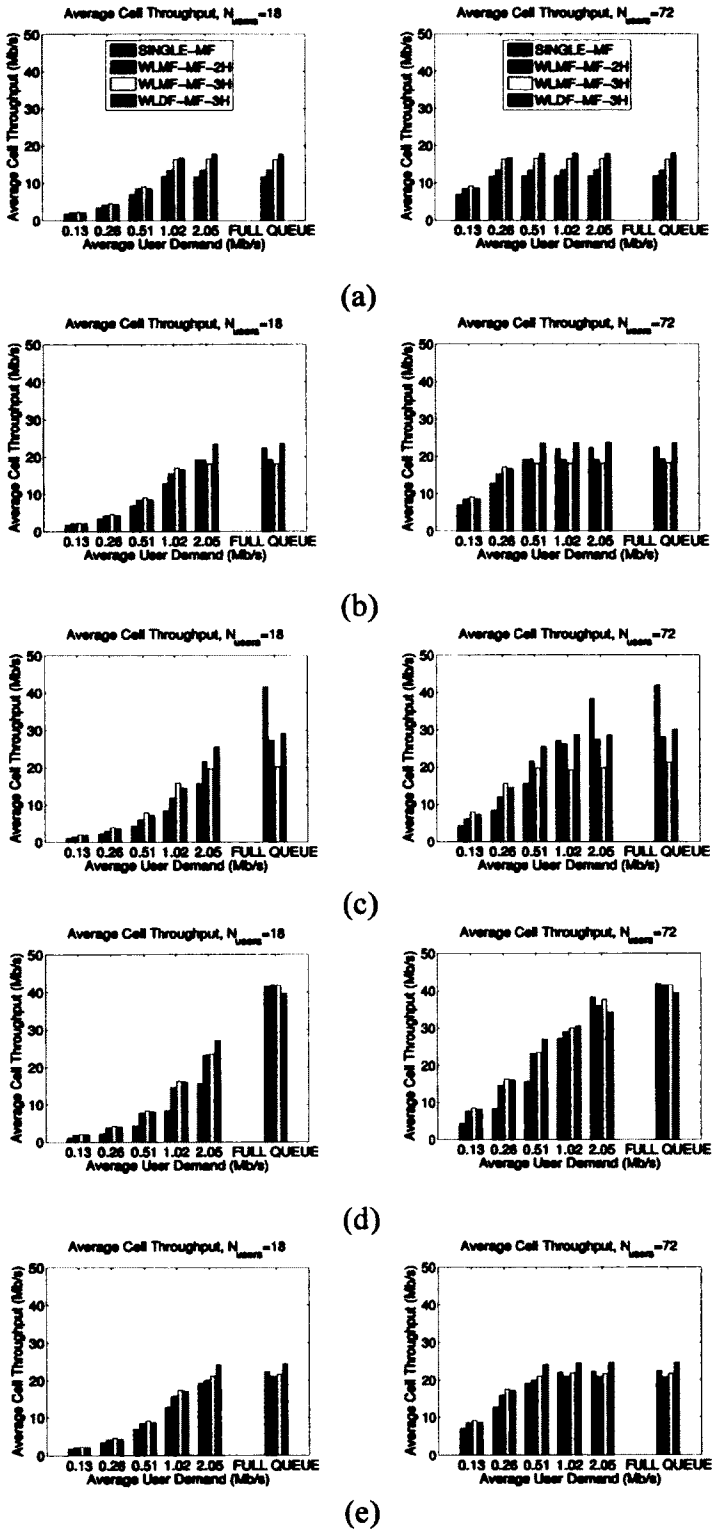


Figure 4-7 Average cell throughput using different scheduling utilities: (a) MF, (b) PF, (c) MT, (d) E2EMT, and (e) E2EPF. Two numbers of users considered: 18 (left) and 72 (right).

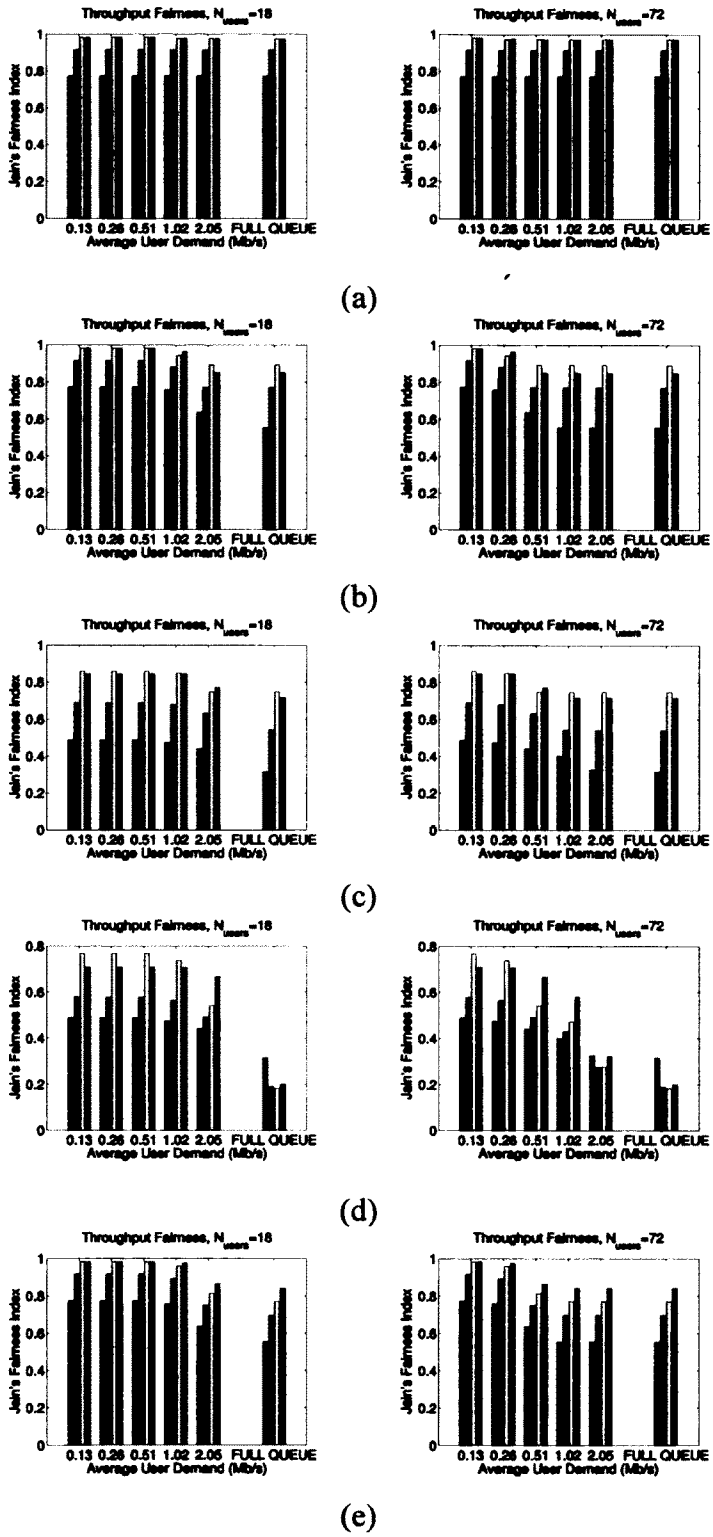


Figure 4-8 Average throughput fairness using different scheduling utilities: (a) MF, (b) PF, (c) MT, (d) E2EMT, and (e) E2EPF. Two numbers of users considered: 18 (left) and 72 (right).

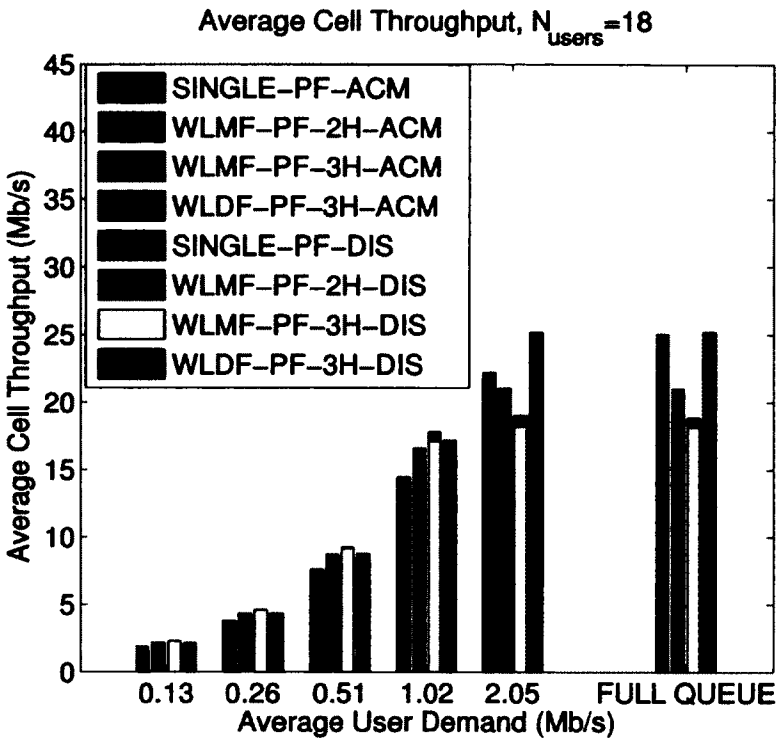


Figure 4-9 Effect of permutation type on throughput performance, contagious (ASP) vs. Distributed (DIS).

Chapter 5

Tunnel Based Scheduling Algorithm

5.1 Introduction

Our results in previous chapters indicate the importance of IRRR in improving the utilization of radio resources. Static IRRR has been discussed earlier in section 3.1 for 3-hop and 2-hop relay networks. It has further been studied for two-hop systems in many papers in the literature such as [67]-[69]. While static methods of radio partition and reuse schemes that were studied in section 3.1 may improve system spectral efficiency, dynamic methods would improve the performance further by adapting the allocation according to changes in traffic and/or channel conditions.

The dynamic allocation problem for multi-hop relay networks was found to be NP hard [9] even without IRRR capability. With a large number of users and a large number of resource units available, search algorithms may not converge to the optimal solution within a single scheduling interval. Furthermore, the overhead of feedback and management massages would be a burden for increasing the hop count beyond two. These factors drift the trend in the literature toward two-hop relaying only. Different allocation algorithms devised solely for two-hop relay networks, such as [5], [8], and [53], which provide dynamic fair allocation algorithms, or maximize the

throughput as in [6]. Some solutions use a graph theoretic approach to jointly optimize routing and subcarrier allocation, maximizing the throughput as in [7] and assuring that a minimum amount of resources is allocated to each user. They use a heuristic two-hop approach based on a proportional fair algorithm. It should be noted that two-hop based approaches are numerous, and a good literature review of some of these approaches can be found in [81]. These algorithms are implemented for two-hop deployments only and, therefore, they are not suitable for a higher hop count. Although a two-hop deployment would be adequate and more practical for most cases, increasing hop count with improved IRRR can provide better performance in terms of throughput and coverage. Relay operations increase the demand on system resources. Therefore, improving IRRR is essential to improve the efficiency of relay networks.

Algorithms such as [10] and [9] rely on an unrealistic assumption to provide a multi-hop (more than 2-hops) solution as they assume the forwarding of traffic can be handled within a single frame. While it is possible to practically adopt some of these solutions as we had done in the previous chapter, these solutions would still remain unable to employ IRRR.

Although it was not designed for cellular systems, [62] propose a multi-purpose solution for multi-hop relay networks that involves joint routing and scheduling. It also involves odd-even frame partitioning to satisfy the half-duplex constraint, implying that the father relay cannot be simultaneously active. Furthermore, it assumes all relay links in the same hop level to be orthogonal, making simultaneous communication possible for links of the same hop level; that is IRRR capability

invoked by a routing scheme. A link cannot be selected in a route if it interferes with the adjacent routes. Therefore, if any two relays are more than two hops apart, they can transmit simultaneously. The main drawback of this strategy is that the routing method they propose does not guarantee better overall throughput or fairness performance. In fact, their method may easily lead to longer routes and increased bottlenecks.

In addition, combining the routing problem with the scheduling problem increases the complexity, which in turn increases the overhead and delays. Typically, fixed relays are well-established and it is unnecessary to change their routing tables with every scheduling instance. Some recovery mechanism may certainly be implemented to reroute disconnected relays or to enhance load balancing, but this should not be applied with every scheduling cycle.

5.2 Contributions

1. We have developed a new scalable resource scheduling algorithm referred to as Tunnel Based Scheduling Algorithm (TBSA) devised for relay enhanced cellular OFDMA networks. It is designed not only for two-hop networks, as in the majority of the proposed solutions in the literature, but it can also accept any number of hops. TBSA. has the following features:

    a. It utilizes the end-to-end tunneling concept to reduce the complexity of the scheduling algorithm. We envisioned tunnel utilization metrics to keep track of the unutilized resources. This improves the utilization while minimizing the need for an exhaustive search for available resources.

    b. It enables IRRR to enhance the throughput and to reduce algorithm complexity.

2. We developed a novel pre-allocation procedure to adapt zone and group allocation, which allows for throughput enhancement and load balancing, unlike [9] or [10], which assign resources based on the number of users only. As stated in [11], system throughput cannot be optimized if the system isn't traffic or queue aware. Therefore, load balancing is critical to optimize the resource utilization; that is, our system adapts resource allocation to match traffic conditions, as well as channel conditions. The resulting algorithm is given the name Adaptive-TBSA (A-TBSA).

5.3 System description

We consider a cell surrounded by 18 other cells in an OFDMA network. A RS, the BS, or a sector of the BS will be referred to as Access Nodes (AN) if it serves the UT's that are distributed all over the cell area. Furthermore, an Access Entity (AE) is

formed by one AN or a group of AN's that share a common pool of resources. An end-to end Tunnel is established between the BS and the AE to aggregate its access traffic.

For AE a, a route vector, $\Re(a)$, is defined as an ordered set of the Relay Links (RL) that constitute the path of the tunnel. We can write the route vector as $\Re(a) = \{\mathfrak{F}^1(a) = BS, \dots, \mathfrak{F}^h(a), \dots, \mathfrak{F}^{h_a-1}(a)\}$, where $\mathfrak{F}^h(a)$ is the h^{th} hop father of RS a counted from the BS. If h=1, the father node is the BS, otherwise it is a relay at hop count $h-1$ from the BS, where $h \in \{1, \dots, h_a - 1\}$, and h_a is the hop count of RS a access links. For each AE a, the scheduler is required to find the variables $\{x_{\mathfrak{F}^h(a),f,t}(\tau(h))\}$ and $x_{a,f,t}(\tau(h_a))$, where $x_{\mathfrak{F}^h(a),f,t}(\tau(h))$ is the relayed traffic allocation variable of slot (f,t) over the h^{th} hop link directed to AE a applied on time frame $(\tau(h))$ and, given by,

$$x_{\mathfrak{F}^h(a),f,t}(\tau(h)) = \begin{cases} 1 & \textit{If RS a is allocated slot } (f,t) \\ & \textit{for its } h^{th} \textit{ hop relay link} \\ & \textit{applied on time frame } (\tau(h)) \\ 0 & \textit{Otherwise} \end{cases} \tag{5.1}$$

Understanding timing notations is critical in understanding the algorithm. We have denoted the argument of the allocation variables by $\tau(h)$, which includes the link hop count, h, to indicate that the application time $\tau(h)$ is related to the BS transmission time (scheduling time) by the equation

$$\tau(h) = v + (h-1)t_\tau \tag{5.2}$$

where v is the scheduling time (time of transmission by the BS) and t_r is the relaying delay normalized by frame duration and rounded to the smallest integer larger than or equal to the normalized relaying delay. It is worth mentioning that the application time $\tau(h)$, is the time at which the traffic is received by the receiver, which is the h^{th} father node in this case. We will assume t_r equals 1, and hence $\tau(h) = v + h - 1$, meaning that the RS will be able to relay the traffic it receives within the next frame after the frame during which it receives the relayed traffic. Therefore, we have

$$x_{\Im^h(a),f,t}(\tau(h)) = x_{\Im^h(a),f,t}(v + h - 1) \tag{5.3}$$

On the other hand, $x_{a,f,t}(\tau(h_a))$ is the allocation variable in the access zone which is used to forward the traffic to the destination UT. At v, we find $x_{a,f,t}(\tau(h_a))$ and $x_{\Im^h(a),f,t}(\tau(h)), \forall a, f, t.$

Channel quality in bits per slot of each flow is assumed to be known at the BS from which it can estimate the number of slots needed for transmission. We designate $q_i(v - h_a)$ as the channel quality estimate received by the BS before scheduling time v, which may be used as the access link channel quality indicator. Alternatively, we may use the moving average estimator of channel, $\tilde{q}_i(v)$ using the following relation:

$$\tilde{q}_i(v) = \left(1 - \frac{1}{tc}\right)\tilde{q}_i(v-1) + \frac{1}{tc}q_i(v - h_a), \forall i \in \mathcal{U}(a) \tag{5.4}$$

Let $x_{i,f,t}(\tau(h_a))$ be slot (f, t) assignment for user i on the access link scheduled at v and will be applied on $\tau(h_a)$. The total number of slots scheduled for access traffic of user i, $\tilde{x}_i(\tau(h_a))$, can be calculated as

$$\tilde{x}_i(\tau(h_a)) = \left\lceil \frac{D_i(v)}{\tilde{q}_i(v)} \right\rceil \tag{5.5}$$

where $D_i(v)$ is the amount of data in bits scheduled on v for flow i, which is dependent on queue size and available resources.

Assuming that some flows belonging to the tunnel a are scheduled, we can calculate $\tilde{x}_a(\tau + h_a)$, the total number of slots needed for access links associated with AE a, as

$$\tilde{x}_a(\tau(h_a)) = \sum_{\forall i \in U(a)} \tilde{x}_i(\tau(h_a)) \tag{5.6}$$

The number of slots needed for each relay hop link, $\tilde{x}_{\mathfrak{F}^h(a)}(\tau(h))$, along the path of tunnel a can be calculated based on the size of data and relay links quality, as

$$\tilde{x}_{\mathfrak{F}^h(a)}(\tau(h)) = \sum_{\forall a \in \mathfrak{D}(\mathfrak{F}^h(a))} \frac{D_a(v)}{\tilde{q}_{\mathfrak{F}^h(a)}(v)}, h = 1, \dots, h_a - 1 \tag{5.7}$$

where $\tilde{q}_{\mathfrak{F}^h(a)}(v)$ is the estimated channel quality of the h^{th} hop relay link and $D_a(v)$ is total traffic scheduled at v to be transmitted through the tunnel a, which is equal to

$$D_a(v) = \sum_{\forall i \in U(a)} D_i(v) \tag{5.8}$$

To enable reuse capabilities, we propose to group links in a manner that simultaneous transmission is possible. For AE j, we define $g_R(j)$ and $g_A(j)$ as a group in the relay zone and access zone, respectively, to which AE j belongs. We update the number of slots associated with each group by defining $x_{g_R(j),f,t}(\tau(h))$ and

$x_{g_A(j),f,t}(\tau(h_a))$ as the assignment of those two groups and $\tilde{x}_{g_R(j)}(\tau(h))$ and $\tilde{x}_{g_A(j)}(\tau(h_a))$ as the number of slots scheduled in them.

The number of resources available for a link is limited by its group resources. That is, the number of resources allocated to relay links is limited by

$$\tilde{x}_{\mathfrak{F}^h(a)}(\tau(h)) \leq \tilde{x}_{g_R(\mathfrak{F}^h(a))}(\tau(h)), \forall a, h \tag{5.9}$$

Similarly, the groups in access zones limit the access link assignment according to

$$\tilde{x}_a(\tau(h_a)) \leq \tilde{x}_{g_A(a)}(\tau(h_a)) \tag{5.10}$$

In static TBSA, we assume fixed zone assignment and fixed group assignment. It is assumed that the frame is divided into two parts: one for access traffic, called Access Zone (AZ), and the other for relay traffic, called Relay Zone (RZ). Group assignments for both the RZ and AZ are assumed static regardless of traffic or user distribution. The groups are assigned resources in a manner that they do not exceed zone resources. Hence, for RZ we have

$$\sum_{g_r=1}^{N_{g_r}} \tilde{x}_{g_r}(\tau(h)) \leq Z_R(\tau(h)) \tag{5.11}$$

for AZ we have

$$\sum_{g_a=1}^{N_{g_a}} \tilde{x}_{g_a}(\tau(h_a)) \leq Z_A(\tau(h_a)) \tag{5.12}$$

Static resource allocation of groups and zones may limit utilization, since some groups may have a lower load compared to others. Thus, the system is better to adapt such allocation process to cope with the demand to improve utilization and provide load balancing. However, this adaptation process requires all links on all hops to

same time. In the next section, we propose a pre-allocation process to inform all links of their synchronized adaptive schedule at the same time.

5.4 Tunneling Concept

Tunneling in relaying context is a mechanism by which a number of flows are aggregated over a link or multiple links in order to reduce management overhead. According to [82], we can distinguish between two types of tunnels: hop-by-hop tunnels and end-to-end tunnels. In hop-by-hop tunnels, a tunnel is established over one link across which all traffic is aggregated. The aggregation of traffic requires encapsulating the MPDUs of different flows traversing the same tunnel into tunnel packets with a unique ID distinguishing them from other tunneled or non-tunneled packets. If the tunnel is between the BS and the AE, it is classified as end-to-end. Typically tunnels are unidirectional, meaning that the tunnel from the BS to an AE is different from the one that originates from the same AN to the BS. [82] showed that the improvement in MAC efficiency can exceed 80%, especially if the number of flows is high. The benefit from tunneling depends heavily on the number of users that are supported. As the number of users decreases, we should expect the gain from the tunnel to decrease to the point that tunnels may cause a reduction in throughput.

Besides the benefit of tunnels in improving MAC efficiency, they can simplify the scheduling problem. *With **end-to-end** tunnels, we can reduce the complexity of the scheduling problem by aggregating the flows into one tunnel to compete with other tunnels on relay resources.* One of the difficulties of the allocation problem in multi-hop relay networks is that for a single flow, *multiple relay links* should be allocated

adequate resources to forward *the same data* at *different time frames. The same resources may be used for other competing flows with different sets of links.* This requires the scheduler to check for previously scheduled flows that may have resources scheduled to them on the current or future frame time. This process should be carried out for every flow on every link on each hop for the next h_{max} time frames. This could be tedious and time consuming, especially for a large number of users. We reduced the complexity of the problem by reducing the number of competing entities from N_u, the number of users, to N_t, the number of tunnels. Moreover, we reduced the complexity further by correcting the route. In the optimal case, the route is part of the allocation problem.

Ignoring the overhead cost, we should expect degradation in throughput compared to the optimal case since the feedback and addressing overhead associated with optimal case are very large. We expect the reduction in throughput of TBSA to be minimal if overhead is taken in consideration.

Figure 5.1-a depicts an example of the end-to-end tunneling principle where we assume tunnels between the BS and each RS. In our linear network example, we have two tunnels connecting the BS to RS_1 and RS_2. Any data received by the BS which belongs to one of the relayed UT's will be aggregated in a tunnel associated with the serving AN. For example, if the data received at the BS belongs to UT_B, then it will traverse tunnel RS_1, while for UT_C and UT_D, their traffic will be aggregated into tunnel RS_2.

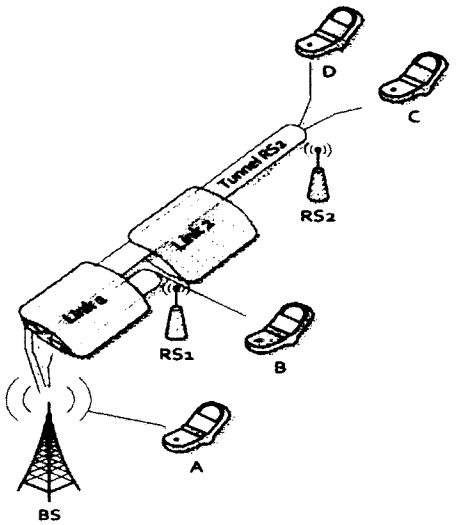


Figure 5.1 Schematic Diagram of TBSA Example

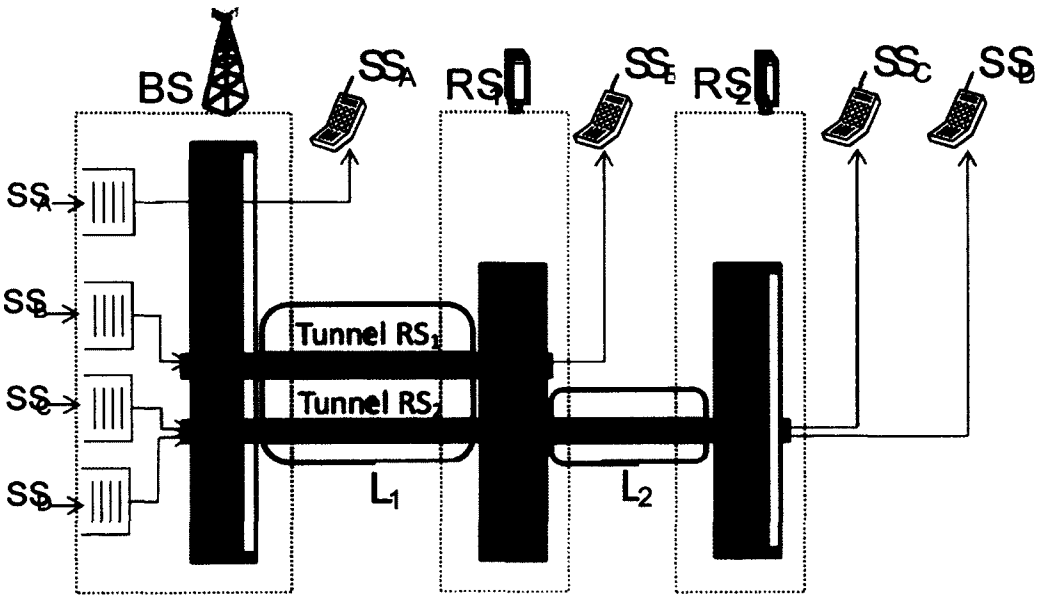


Figure 5.2 Block Diagram of TBSA Example

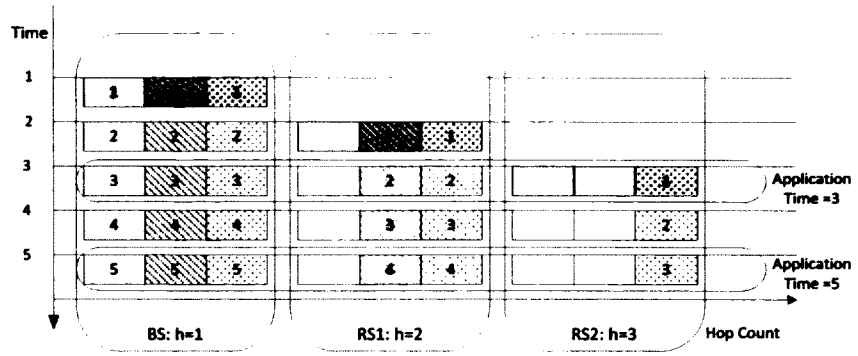


Figure 5.3 Timing Diagram of TBSA Example

The key factor in this allocation paradigm is the implementation of tunnel utilization metrics for access traffic assignment called tunnel access metric, $\mu_{a_A}(\tau)$, which computes the available radio resources for access traffic. It is initialized by the access group allocation and calculated as the remaining resource assigned to the access group not utilized by the designated tunnel. This can be written as

$$\mu_{a_A}(v) = \tilde{x}_{g_A(a)}\left(\tau(h_a)\right) - \tilde{x}_a\left(\tau(h_a)\right) \tag{5.13}$$

Similarly, we can denote the tunnel relay utilization metric by $\mu_{a_R}(v)$, which calculates the amount of data that can be added to the tunnel without exceeding relay group assignments at all links. We can define the relay utilization metric of tunnel a as the difference between the relay group assignment $\tilde{x}_{g_R\left(\mathfrak{F}^h(a)\right)}$ and the relay link assignment $\tilde{x}_{\mathfrak{F}^h(a)}$ on every hop, i.e.

$$\mu_{a_R}(v) = \min_{h=1:h_a-1}\left(\left(\tilde{x}_{g_R\left(\mathfrak{F}^h(a)\right)}\left(\tau(h)\right) - \tilde{x}_{\mathfrak{F}^h(a)}\left(\tau(h)\right)\right)\tilde{q}_{\mathfrak{F}^h(a)}\right) \tag{5.14}$$

5.5 Static TBSA Scheduling Algorithm

We will rely on flow utility maximization to select the nominated flows. Then, the required slots on each link are assigned. The scheduler makes sure that it has enough resources to schedule the flow. If there are insufficient resources, either the number of scheduled packets is reduced or another flow is selected. Otherwise, the flow is rejected and the scheduler selects the next best flow. If the flow is selected, the scheduler finds the next best flow and so on.

The scheduled flow denoted by $i^* \in a^*$ is assigned a number of resources equal to $\hat{x}_{i^*}(\tau(h_{a^*}))$, which is dependent on $D_{i^*}(v)$. The amount of data that can be transmitted to i^*. $D_{i^*}(v)$ is limited by

1. Queue length of flow i^*, $Q_{i^*}(v)$.
2. The unutilized access resources which are the sum of unutilized access group resources, $\mu_{a_A}(v)$.
3. The unutilized relay resources which equal the relay utilization metrics for the tunnel, $\mu_{a_R}(v)$.

Therefore, we can depict the scheduled traffic for flow i^* as

$$D_{i^*}(v) = \text{minp}\left(Q_{i^*}(v), min\left(q_i(v)\mu_{a_A}(v), \mu_{a_R}(v)\right)\right) \tag{5.15}$$

where "minp(Q_i, β)" is packet-wise minimum, which is equal to the maximum number of bits of queued packets of flow i that can be scheduled without exceeding β

bits. After the allocation procedure, we will update the utility metrics. Algorithm 5.1 shows the pseudo-code for our Static TBSA (S-TBSA) algorithm.

One of the advantages of our system is its linear complexity. The algorithm is only required to choose the flow with the highest PF utility. The process by which we update relay utility metrics after each flow assignment consumes the time in the algorithm. The process is used to perform $h_j$calculations to compute each relay metric of tunnel j. If we have N_t tunnel, the complexity of S-TBSA is of order $O(N_i N_t h_{max})$.

To better understand the TBSA algorithm, we described the algorithm for the example presented in Figure 5.1-a. Figure 5.1-b depicts a timing diagram of the simple example of a frame transmission operation. A frame transmitted by the BS contains three allocations for traffic belonging to one-hop UT's, two-hop UT's or three-hop UT's. This frame is first received by the one-hop UT's and the one-hop RS's; in this case $h = 1$ implies that the father node is the BS. The one-hop RS's transmit the received traffic to its access UT's or to the next hop RS's; in this case $h = 2$. The third hop link connects the two-hop RS's to its access users ($h = 3$). According to the diagram, the traffic intended for three-hop UT's require three frames to arrive to the destination. Notice that while the three-hop traffic is scheduled and transmitted by the BS at time frame 1, for example, it will not be received before time frame 3.

At time 5, we notice that the frame scheduled at 5 competes with relayed traffic scheduled at 4, and that both compete with the 3rd hop traffic scheduled at 3.The traffic of three-hops UT's scheduled at 3 was also competing with other traffic scheduled at 1. This would imply that when the scheduler operates on the 3-hop UT's, it will check

the traffic scheduled at time 1 (for the three-hop UT's) and the ones after till it reaches time frame 5. The current and next two frames would also need to be scheduled. When competition between different hop allocations occurs on resources, the scheduler does not have control of the previously transmitted allocations. Therefore, before a new allocation, the scheduler checks those previous allocations to avoid interference.

To understand the utilization metrics, we will assume that relay links have already been assigned resources. Moreover, we will assume that each relay has already been assigned resources for its access traffic. Therefore, each relay link has been pre-assigned its resources. We envision tunnel utilization metrics to compute the maximum throughput that can be achieved for tunnels traversing those links. For example, if the link from the BS to RS_1 can accept 2Mbps while the second link (from RS_1 to RS_2) can accept 1Mbps, then the RS_2 and RS_1 relay tunnel metrics are 1Mbps and 2Mbps respectively. The access tunnel metric computes the maximum throughput that can be achieved by utilizing access links. If the resources allocated for RS_1 and RS_2 access traffic allow for only 1.5Mbps, which represent the access utilization metric of tunnels RS_1 and RS_2, then we can compare both the relay and access utilization metrics and choose the minimum. Thus, we will have 1.5Mbps as the utilization metric for RS_1 tunnel and 1Mbps for RS_2 tunnel.

We adopt a flow based allocation where a PF metric is computed for each flow. If the BS has packets to send to UT B, which has the highest PF metric, the BS will assign some or all of the resources that are available for RS_1 tunnel. Thus, the BS will check RS_1 tunnel metrics. Resources are assigned such that either there are no more

Algorithm 5.1 Static TBSA

Static TBSA Algorithm

This procedure is activated every scheduling instant v. It assumes the assignment of access groups $\{\tilde{x}_{g_A}\}$ and relay groups $\{\tilde{x}_{g_R}\}$ is known. It is also assumes that a number of flows N_t are waiting to be scheduled. For each flow, $Q_l(v)$ and $q_l(v)$ are known. All other variables are initialized with zeros.

$\mu_{a_A}(\tau) = \tilde{x}_{g_{A(a)}}(\tau)$

$\mu_{a_R}(\tau) = \tilde{x}_{g_{R(a)}}(\tau)$

$Count = N_t$

While $\mu_{a_A}(\tau) > 0 \;\&\; \mu_{a_R}(\tau) > 0 \;\&\; Count > 0$

Count=*Count*-1;

$i^* = \max_{\forall l \in}(u_l)$

$D_{l^*}(v) = \min p\left(Q_{l^*}(v), q_{l^*}(v)\mu_{(a_A^*}(v), \mu_{a_R^*}(v)\right)$

$D_{a^*}(v) = D_{a^*}(v) + D_{l^*}(v)$

$\tilde{x}_{a^*}(\tau(h_{a^*})) = \tilde{x}_{a^*}(\tau(h_a)) + \tilde{x}_{l^*}(\tau(h_{a^*}))$

$\tilde{x}_{\mathfrak{F}^h(a^*)}(\tau(h)) = \tilde{x}_{\mathfrak{F}^h(a^*)}(\tau(h)) + \frac{D_a(\tau)}{\tilde{q}_{\mathfrak{F}^h(a^*)}(\tau)};\; h = 1, \ldots, h_{a^*} - 1$

$\tilde{x}_{g_R(\mathfrak{F}^h(a))}(\tau(h)) = \max\left(\tilde{x}_{g_R(\mathfrak{F}^h(a))}(\tau(h)), \tilde{x}_{\mathfrak{F}^h(a)}(\tau(h))\right)$

$\tilde{x}_{g_A(a^*)}(\tau(h_{a^*})) = \max(\tilde{x}_{g_A(a^*)}(\tau(h_{a^*})), \tilde{x}_a(\tau(h_{a^*})))$

$\mu_{a_A}(v) = \tilde{x}_{g_A(a)}(\tau(h_a)) - \tilde{x}_a(\tau(h_a))$

$\forall tunnel\ j\ Do$

$\mu_{j_R}(v) = \min_{h=1:h_j-1}\left(\tilde{q}_{\mathfrak{F}^h(a)}\left(\tilde{x}_{g_R(\mathfrak{F}^h(a))}(\tau(h)) - \tilde{y}_{\mathfrak{F}^h(a)}(\tau(h))\right) : \mathfrak{F}^h(j) = \mathfrak{F}^h(a)\right)$

END Do

END While

resources, i.e., tunnel metric becomes zero or there are no more packets in the UT_B queue. If UT_B consumes some or all tunnel RS_1 resources, the utilization metrics of all tunnels using them are updated. For example, if we schedule 1.5Mbps for UT_B and accordingly the RS_1 tunnel consumes all possible resources, the relay utilization metrics of RS_1 as well as RS_2 are updated, since the available resources for the first relay link are not 2Mbps anymore, but reduced to 0.5 Mbps. Therefore, the relay utilization metric becomes zero for RS_1 tunnel and 0.5 Mbps for RS_2 tunnel.

5.6 Adaptive grouping and zone assignment

To improve the utilization of the system, we propose to adapt the size of allocation groups and zones. Performing such an adaptation requires a synchronized allocation on all hops at the same time. This also requires what we referred to as a pre-allocation process. This process will adaptively change the groups and zones, which activates the new zone's size at the same time on all hops to avoid overlapping assignments on different hops. This implies that the BS is required to decide on the new assignments on all hops simultaneously, which will be activated after a delay, or t_z, called new zone time given by the inequality

$$t_z \geq h_{max} t_r \geq h_{max} \tag{5.16}$$

If $\tau(h)$ is the application time, the pre-allocation time will be $б = \tau(h) - h_{max} + 1$. The duty of the pre-allocation process is to perform the preliminary scheduling on time frame б for all links on all hops to be applied on time $\tau(h)$.

We will define the variables $y_{\mathfrak{F}^h(a),f,t}(\tau)$, $y_{a,f,t}(\tau)$ as the allocation variable of slot (f,t) pre-allocated on б , transmitted by BS at $v(h)$, applied on τ for the h^{th} relay link and for access link of AE a ($h = h_a$ for access link), respectively. $\tilde{\mathfrak{X}}_{\mathfrak{F}^h(a)}(\tau)$ and $\tilde{\mathfrak{X}}_{\mathfrak{a}}(\tau)$ are their total numbers. The relationship between the three time instances is as follows:

$$б = \tau - h_{max} + 1, \text{for all links} \tag{5.17}$$

$$v(h) = \tau - h + 1 = б + h_{max} - h, \text{for the } h^{\text{th}} \text{ hop link} \tag{5.18}$$

In the Static TBSA scheduler presented in the previous sections, we noticed that the transmission time is fixed, while the application time varies according to hop

count. In contrast, the pre-scheduler requires fixed application time,τ, while the transmission time, $v(h)$, can vary based on hop count.

Flow utility optimization is proposed as a method of the UT's flow selection method. At transmission time v, a flow i will receive a new traffic with size equal to $B_i(v)$, assuming a queue for each flow storing the waiting traffic of size $Q_i(v)$ and the scheduled traffic of flow i is to be denoted by $D_i(v)$. Those variables are estimated for pre-allocation purpose to guarantee an interference-free IRRR operation. Therefore, we will use the variables: $\tilde{B}_i(б)$, $\tilde{Q}_i(б)$ and $\tilde{D}_i(б)$ for estimating $B_i(v)$, $Q_i(v)$ and $D_i(v)$, respectively.

The estimation of $\tilde{B}_i(б)$ is dependent on the type of traffic. This problem has been a topic of research for a long time and is not trivial. We are not concerned with traffic prediction in this thesis. However, it could be considered as an extension to our work. For our purpose, we will rely on a simple moving average:

$$\tilde{B}_i(б) = \begin{cases} B_i(б) & if\ h_i = h_{max} \\ \left(1 - \frac{h_i}{t_c}\right)\tilde{B}_i(б-1) + \frac{h_i}{t_c}B_i(б) & Otherwise \end{cases} \quad (5.19)$$

where $\tilde{D}_i(б)$ is the data size pre-allocated to flow i, which is an estimation of the actual data size to be allocated during the scheduling process.

The queue size for flow i will be updated according to

$$Q_i(б) = Q_i(б-1) + B_i(б-1) - D_i(б-1) \quad (5.20)$$

The pre-allocation estimation of queue size at time v is proposed to be

$$\tilde{Q}_i(б) = Q_i(б) + \sum_{б'=б-h_{max}+h_a+1}^{б-1} \left(\tilde{B}_i(б') - \tilde{D}_i(б')\right) \tag{5.21}$$

We estimate the access link quality by using the moving average estimator $\tilde{q}_i(б)$ given by

$$\tilde{q}_i(б) = \left(1 - \frac{1}{t_c}\right)\tilde{q}_i(б-1) + \frac{1}{t_c} q_i(б - h_a) \tag{5.22}$$

The first step of the pre-allocator is to make sure that all relayed traffic is forwarded to the next hop node. The next hop node may be an RS or a UT. For an RS a on the h^{th} hop, we define the pre-allocation variable as $y_{\mathfrak{F}^h(a),f,t}(б + h - 1)$. We can further define $\tilde{y}_{\mathfrak{F}^h(a)}(б + h - 1)$ as the number of pre-allocated slots, $h = 1, \ldots, h_a - 1$.. For the ultimate destination, we define $y_{i,f,t}(\tau)$ and $\tilde{y}_i(\tau)$ as the indicator variable and the number of pre-allocated slots of the access link of flow i, $i \in \mathcal{U}(a)$, respectively. We can also define $\tilde{y}_a(\tau)$ as the number of pre-allocated slots for the access traffic of RS a which is the sum of all flows traffic it serves:

$$\tilde{y}_a(\tau) = \sum_{\forall i \in \mathcal{U}(a)} \tilde{y}_i(\tau) \tag{5.23}$$

After performing the pre-allocation for relayed traffic and ensuring that radio resources are available for accommodating new traffic, we can employ a flow utility maximization to schedule new flows. This step will determine $\tilde{y}_{\mathfrak{F}^1(a)}(\tau)$ for flows served by RS's or $\tilde{y}_{BS}(\tau)$ for flows served by BS.

The data to be pre-allocated for flow i is denoted by $\tilde{D}_i(б)$ and is limited by available slots as well as the queued data as follows:

$$\tilde{D}_i(б) = \text{minp}\Big(\tilde{Q}_i(б), \Big(\min\Big(\tilde{q}_i(б)\mu_z(v(h_{a^\cdot})) + \mu_{a_A}(v(h_a)), \mu_{a_R}(v(h_a)) + \min_{h=1:h_a-1}(\mu_z(б+h-1)\tilde{q}_{\mathfrak{F}^h(a)}(б))\Big)\Big)\Big) \tag{5.24}$$

where μ_z is the utilization metric for all zones which equals the number of resources that have not been utilized yet.

The estimated total traffic for RS a denoted by $\tilde{D}_a(б)$ is given by:

$$\tilde{D}_a(б) = \sum_{\forall i \in U(a)} \tilde{D}_i(б) \tag{5.25}$$

Accordingly the pre-allocated resources of the first hop can be written as $\tilde{\mathcal{Y}}_{\mathfrak{F}^1(a)}(б) = \left\lceil \frac{\tilde{D}_a(б)}{\tilde{q}_{\mathfrak{F}^1(a)}(б)} \right\rceil$ for the relayed traffic, and $\tilde{\mathcal{Y}}_i(б) = \left\lceil \frac{\tilde{D}_i(б)}{\tilde{q}_i(б)} \right\rceil$ for flows served directly by BS. The relay links are incrementally updated according to the new scheduled traffic

$$\tilde{\mathcal{Y}}_{\mathfrak{F}^h(a)}(б+h-1) = \tilde{\mathcal{Y}}_{\mathfrak{F}^h(a)}(б+h-1) + \left\lceil \frac{\tilde{D}_i(б)}{\tilde{q}_{\mathfrak{F}^h(a)}(б)} \right\rceil \tag{5.26}$$

Where $h = 1, \ldots, h_a - 1$. For the access relay, the pre-allocator should determine the required resources for access traffic using the equation

$$\tilde{\mathcal{Y}}_i(\tau) = \left\lceil \frac{\tilde{D}_i(б)}{\tilde{q}_i(б)} \right\rceil \tag{5.27}$$

for the i^{th} flow and equation 5.21 for the total access traffic of RS a.

IRRR capability is supported by grouping the relays based on their interference where we define $g_R(j)$ and $g_A(j)$ respectively as groups in the RZ and the AZ to

which RS j belongs. We further define $x_{g_r(j),f,t}(б+h-1)$ and $x_{g_a(j),f,t}(\tau)$ as the assignments of those two groups, and $\hat{x}_{g_r(j)}(б+h-1)$ and $\hat{x}_{g_a(j)}(\tau)$ as their number of scheduled slots which are, respectively, given by

$$\hat{x}_{g_r(\mathfrak{F}^h(a))}(б+h-1) = \max_{\mathfrak{F}^h(a)\in g_r(\mathfrak{F}^h(a))} \hat{y}_{\mathfrak{F}^h(a)}(б+h-1), \forall a,h \tag{5.28}$$

$$\hat{x}_{g_A(a)}(\tau) = \max_{a\in g_A(a)} \hat{y}_a(\tau), \forall a,h \tag{5.29}$$

The zones are updated according to

$$Z_R(б+h-1) = \sum_{g_r=1}^{N_{g_r}} \hat{x}_{g_r}(б+h-1), h = 1{:}\,h_{max}-1 \tag{5.30}$$

$$Z_A(\tau) = \sum_{g_a=1}^{N_{g_a}} \hat{x}_{g_a}(\tau) \tag{5.31}$$

Note that $Z_A(\tau)+Z_R(\tau)$ should be less than the total number of slots.

Algorithm 5.2 Adaptive TBSA

Pre-Allocation Procedure

$Count = N_l$

While $\mu_{a_A}(\tau) > 0 \,\&\, \mu_{a_R}(\tau) > 0 \,\&\, Count > 0$

Count=Count-1;

$i^* = \max_{\forall l \in a}(u_l)$

$\tilde{D}_l(б) = \min p\left(\tilde{Q}_l(б), \left(\min\left(\tilde{q}_l(б)\mu_z(v(h_{a^*})) + \mu_{a_A}(v(h_a)), \mu_{a_R}(v(h_a)\right) + min_{h=1:h_a-1}(\mu_Z(б + h - 1)\tilde{q}_{\mathfrak{F}^h(a)}(б))\right)\right)$

$\tilde{D}_a(б) = \tilde{D}_a(б) + \tilde{D}_l(б)$

$\tilde{y}_l(\tau) = \frac{\tilde{D}_l(б)}{\tilde{q}_l(б)}$

$\tilde{y}_a(\tau) = \tilde{y}_a(\tau) + \tilde{y}_l(\tau)$

$\tilde{x}_{g_A(a)}(\tau) = \max(\tilde{x}_{g_A(a)}(\tau), \tilde{y}_a(\tau))$

$\tilde{y}_{\mathfrak{F}^h(a)}(б + h - 1) = \tilde{y}_{\mathfrak{F}^h(a)}(б + h - 1) + \left\lceil \frac{\tilde{D}_l(б)}{\tilde{q}_{\mathfrak{F}^h(a)}(б)} \right\rceil, h = 1, \ldots, h_a - 1$

$\tilde{x}_{g_R(\mathfrak{F}^h(a))}(б + h - 1) = \max\left(\tilde{x}_{g_R(\mathfrak{F}^h(a))}(б + h - 1), \tilde{y}_{\mathfrak{F}^h(a)}(б + h - 1)\right)$

$\mu_{a_A}(v(h_a)) = \tilde{x}_{g_A(a)}(v(h_a)) - \tilde{x}_a(v(h_a))$

$\forall tunnel\ j\ Do$

$$\mu_{j_R}(v(h_a)) = \min_{h=1:h_a-1}\left(\tilde{q}_{\mathfrak{F}^h(a)}\left(\tilde{x}_{g_R(\mathfrak{F}^h(a))}(б + h - 1) - \tilde{y}_{\mathfrak{F}^h(a)}(б + h - 1)\right) : \mathfrak{F}^h(j) = \mathfrak{F}^h(a)\right)$$

END Do

END While

$Z_R(б + h - 1) = \sum_{g_r=1}^{N_{g_r}} \tilde{x}_{g_r}(б + h - 1), h = 1, \ldots, h_{max} - 1$

$Z_A(\tau) = \sum_{g_a=1}^{N_{g_a}} \tilde{x}_{g_a}(\tau)$

$\mu_z(б + h - 1) = \mu_z(б + h - 1) - Z_R(б + h - 1) - Z_A(б + h - 1)$

Once the process finishes pre-allocating the relayed traffic, it starts pre-allocating new traffic by means of flows utility maximization, as we mentioned earlier. This procedure continues until all resources are consumed or until there is no

more traffic waiting for transmission. At the end of the pre-allocation cycle, the scheduler will be informed by all group assignments, which will be used as input for the static TBSA algorithm aforementioned.

The same of order of complexity is maintained by A-TBSA, the algorithm as S-TBSA, which is found to be $O(N_i N_t h_{max})$. Such linear complexity makes the algorithm scalable to a higher number of hops and a larger number of users.

5.7 Link Grouping

TBSA algorithm supports reuse capability based on grouping the links so that links in a group can be activated simultaneously. A zone-based allocation is assumed, which requires two zones to forward traffic. The first is an access zone for accessing traffic from RS or the BS to the directly connected UT, and the second is a relaying zone, which is used to transmit the relayed traffic.

In the access zone, the links that belong to access stations will be grouped into non-overlapping groups so that the interference can be tolerated between the links in a group. Note that an entry of a group may be access links that are connected to either the BS, a single RS, or a group of RS's. Regardless of the option selected, a tunnel is formed for each entry in an access group. The complexity of TBSA algorithms in this case will be of order $O(N_i^2 h_{max}) = O(N_i^2)$, assuming a limited number of hops.

To reduce the complexity, we can group the access links based on the location and interference conditions of UT's. For example, a single tunnel may be formed to

serve a number of UT's served by a sectorized antenna at the BS or an RS. We can also group the UT's of more than one RS into one access group. In the first case, we formed multiple tunnels to a single RS and in the latter, we formed a single tunnel for multiple RS's. In this case, the complexity of TBSA algorithms will be of order $O(N_i N_t h_{max})$ with a limited number of hops and limited number of tunnels. The algorithm complexity will be of order $O(N_i)$.

If Tunnel $a \in g_A$, then $j \in g_A$ if $SIR_{i \leftarrow j} > \widetilde{g_A} SIR_{min}, \forall i \in \mathcal{U}(a), \forall a, j \in g_A$. This implies that the signal to interference ratio between the transmissions of two tunnels in a group is lower bounded by SIR_{min} which is given by

$$SIR_{i \leftarrow j} = \frac{P(j)\mathfrak{h}(i \leftarrow j)}{P(a)\mathfrak{h}(i \leftarrow a)} \tag{5.32}$$

where $P(j)$ is the transmitted power of the RS of tunnel j and $\mathfrak{h}(i \leftarrow j)$ is the channel gain from tunnel j to UT i, $\forall i \in \mathcal{U}(a), \forall a$. This bound is increased as the number of elements in the group increases by multiplying the threshold by $\widetilde{g_A}$, the number of elements in g_A, to account for the increased interference.

Relay links are grouped in relay zones based on interference and half-duplex constraints. The entries of relay groups may be a single RS or a group of RS's sharing the same link. The grouping is done in a manner that the receiving RS (or multiple RS's) can receive the relayed traffic at the same time as other members of the same group. We will assign the name Relay group Entry or RE to refer to the RS or RS's at

reentry. Relay links in the R-zone can be joined in a group and should satisfy the following conditions:

1. If $RE\ a \in g_R$, then $RE\ j \in g_R$ if $SIR_{a \leftarrow \mathfrak{F}(j)} > \widetilde{g_R} SIR_{min}, \forall a, j \in g_R$. That is, the signal to interference ratio between the transmissions received by two RE's in a group is lower bounded by SIR_{min} which is given by

$$SIR_{\mathfrak{F}(j) \to a} = \frac{P(\mathfrak{F}(j))\mathfrak{h}(a \leftarrow \mathfrak{F}(j))}{P(\mathfrak{F}(a))\mathfrak{h}(a \leftarrow \mathfrak{F}(a))} \tag{5.33}$$

where $P(\mathfrak{F}(j))$ is the transmitted power of the father of j and $\mathfrak{h}(a \leftarrow \mathfrak{F}(j))$ is the channel gain from the father node of RE j to RE a. This bound is increased as the number of elements in the group increase to account for the increased interference.

2. If $RE\ a \in g_R$, then $RE\ j \in g_R$ if $j \neq \mathfrak{F}(a) \& j \neq \mathfrak{D}(a), \forall j, a \in g_R$. The direct father of the node a, $\mathfrak{F}(a)$, and the direct child, $\mathfrak{D}(a)$, are not allowed to join the group due to the half duplex constraint.

With the above conditions, we can form a number of groups. These groups may be overlapping, i.e., the same RE may be present in more than one group. We will need to select a non-overlapping set of groups.

In our work, we will consider two examples of cell layout: one with two-hop relays and the other for a three hop case, following the layout presented in Figure 5.4. We assume a uniform distribution of users. We partition the cell such that each

partition contains the same number of users. Each partition will be identified according to its location and its partition number. We use the notation (p, s), where p represents the partition number and $s$represents the sector number. For example, in Figure 5.4-a, the orange partition numbered 4 is indexed by (4,1), where 4 is the partition number and 1 is the sector number.

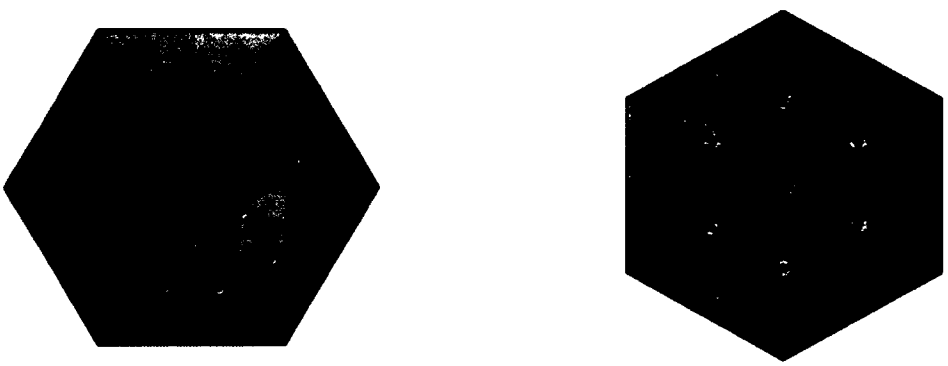
For the three-hop case, six access groups are formed (see Table 5.1-b); each group contains three entries: one of the first tier partitions (one of the six BS sectors), one of the second tier of partitions, and one of the third tier partitions covered by a pair of RS's. There are a total of 18 Tunnels for the 18 partitions depicted in Figure 5.4 (see Table 5.1-a). Tunnels 1, 2, 7, 8, 13 and 14 do not exist in reality, since they belong to the BS coverage. However, they are used for identification purposes for the algorithm to work.

The RE's are formed for the relays according their location. Since we assume that all the relays are capable of serving access traffic, we will use the same tunnel index for RE's. Table 5.1-a shows the list of available RE's. Note that there will be no RE's for the tunnels belonging to the BS. There are six relay groups, with each group containing two entries. One entry is for one of the single-hop relays, while the other is for 3-hop links associated with a pair of adjacent RS's (see Table 5.1-c).

For the two-hop case (Figure 5.4-b), the link grouping is presented in Table 5.2 where we find 6 access groups and 6 relay groups. Each RS will have its own tunnel in addition to 6 BS sectors. Thus we have a total of 12 tunnels. Two entries per access

group and one entry for each relay group, i.e. no relay reuse is assumed (see Table 5.2).

An example of resource allocation for each group is depicted in Figure 5.3. The group allocations are arranged in pairs: group 1 with 2, group 3 with 4 and group 5 with 6. The two groups in a pair share the same time allocation, but different sub-channels. This is because the links in these two groups do not have father/child relationships among each other, and hence they can transmit simultaneously. On the other hand, time division is required between each pair to maintain the half duplex constraint as the links in those pairs violate this condition if they allocated the same time duration. It should be noted that a gap is needed between each pair in the relay zone to allow for switching from transmit state to receive state. This is not the case in the access zone, as all access nodes are in the transmit state.



(a) (b)

Figure 5.4 Cell portioning schemes: (a) three-hop case, and (b) two-hop case.

Partition	Tunnel index (T)	RE
(1,1)	1	
(2,1)	2	
(3,1)	3	3
(4,1)	4	4
(5,1)	5	5
(6,1)	6	6
(1,2)	7	
(2,2)	8	
(3,2)	9	9
(4,2)	10	10
(5,2)	11	11
(6,2)	12	12
(1,3)	13	
(2,3)	14	
(3,3)	15	15
(4,3)	16	16
(5,3)	17	17
(6,3)	18	18

(a)

Access Groups	Member Tunnels		
1	1	9	5
2	2	12	16
3	3	11	13
4	4	8	18
5	5	7	15
6	6	10	14

(b)

Relay Groups	Member RE's	
1	9	5
2	12	16
3	3	11
4	4	18
5	5	15
6	6	10

(c)

Table 5.1 Link Grouping for three-hop case (Figure 5.4-a): (a) Tunnels and RE's indexes for each partition, (b) Access Groups and their members, and (c) Relay groups and their members.

Partition Index	Tunnel Index	RE Index
(1,1)	1	
(2,1)	2	
(3,1)	3	3
(4,1)	4	4
(1,2)	5	
(2,2)	6	
(3,2)	7	7
(4,2)	8	8
(1,3)	9	
(2,3)	10	
(3,3)	11	11
(4,3)	12	12

(a)

Access Groups	Member Tunnels	
1	1	7
2	2	4
3	3	9
4	4	6
5	5	11
6	8	10

(b)

Relay Groups	Member RE's
1	3
2	4
3	7
4	8
5	11
6	12

(c)

Table 5.2 Link grouping for two-hop case (Figure 5.4-b): (a) Tunnels and RE's indexes for each partition, (b) Access Groups and their members, and (c) Relay groups and their members.

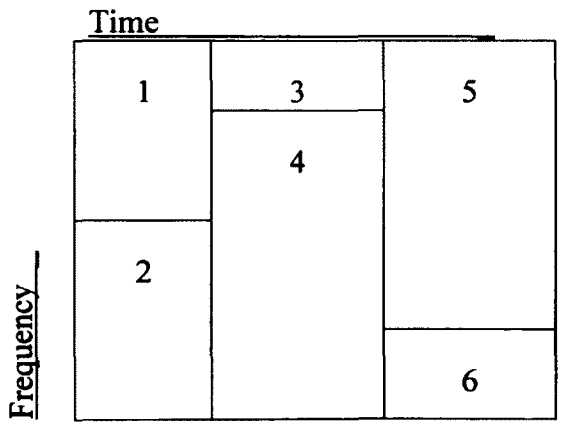


Table 5.3 Group Resource Allocation Example

The TBSA algorithm will work regardless of how link grouping is performed. In fact, link grouping has a significant impact on system performance. Following the rules we stated earlier for group selection, one may optimize the grouping procedure to enhance the performance and match different cell conditions regarding traffic, user distribution, or relay distribution. This problem is not considered in our thesis, but it could be considered as an extension to our work. As an example, we could associate a cost for each RE and we may assume an RE for each tunnel, implying that the BS will have multiple RE's. Access, as well as relay grouping, can be done so that the overall cost of relaying is minimized. The cost at RE a can be calculated as:

$$C_a = C_{A_a} + C_{R_a} \tag{5.34}$$

where C_{A_a} is the cost associated with tunnel a access traffic, if T_{A_a} is the average throughput of RE a access traffic. We can write the cost as

$$C_{A_a} = T_{A_a}/q_{A_a} \tag{5.35}$$

where q_{A_a} is the average access link spectral efficiency in bits/s/Hz.

In a similar manner, we define C_{R_a} as the cost associated with the traffic this RE a relays which is a function of the access traffic of all of its children RE's

$$C_{R_a} = \sum_{j \in \mathfrak{C}(a)} T_{A_j}/q_{R_a} \tag{5.36}$$

where q_{R_a} is the average spectral efficiency of the relay link that connects RE a to its children RE and $\mathfrak{C}(a)$ is the set of children RE of RE a .

The cost of group g_R is the maximum cost of each of its elements

$$C_{g_R} = \max_{\forall a \in g_R} C_{R_a} \tag{5.37}$$

Similarly, the cost of the access groupg_A is given by:

$$C_{g_A} = \max_{\forall a \in g_A} C_{A_a} \tag{5.38}$$

The set of all non-overlapping groups,$\{g_A\}$ and $\{g_R\}$, is selected in a manner that total cost is minimized.

Some groups in the relay zone are time division multiplexed (TDM) due to the half-duplex constraint. Since RS's require to switch between transmission and reception modes, a gap is needed between two groups to allow such a switching if an element in one the groups is a direct father or direct child of a node in the other group. Thus, we can avoid gaps and TDM between two groups if they contain neither direct father nor children elements of each other. The following is an algorithm created to arrange groups in a manner that minimizes switching gaps:

$G = \{g_i\}$; % define G the set of groups containing all groups g_i

$G_s = \phi$; % define G_s an empty ordered set of selected group.

WHILE $G - G_s \neq \phi$

$gs \leftarrow \{G - Gs\}$; % an element is arbitrarily selected from G that has not been selected before.

$G_s \leftarrow gs$; % the group is added to a selected group set.

$i = 1$; % initialize the counter *i*.

$Gs^* = Gs$; % the variable Gs^* is used to temporarily store Gs before % modification.

WHILE$G_s == G_s^* \& \; i \leq i_{max}$ % If G_s is not changed nor we have checked % all groups in *G*.

$i = i + 1$

IF $G(i) \notin G_s$& $G(i) \cap \mathfrak{F}(gs) = \phi$ & $G(i) \cap \mathfrak{D}(gs)$ % half duplex

$Gs \leftarrow G(i)$ % condition is applied. In addition, the group

$gs = G(i)$ % must not be selected more than once.

End IF

END WHILE

IF GS==GS' % if no group satisfied the above condition, TDM and gap is % needed

$GS \leftarrow g_{tdm}$ % a gap g_{tdm} is added in Gs set

$gs = g_{tdm}$

End IF

END WHILE

5.8 Performance Simulation

5.8.1 Simulation Models and Parameters

As per our knowledge, there are no scalable resource allocation algorithms devised for multi-hop relay OFDMA cellular networks that are able to support IRRR. For this reason, we have compared our algorithm with the CPF system presented in chapter [9]. We used the MATLAB simulation environment to test the throughput and fairness performance. The parameters presented in Table 3-1 will be used. The large-scale channel impairments were fixed during each experiment where we performed at least 20 experiments, each of which conform to 50 drops with 200 frames per drop. We assumed a cell that is surrounded by 18 interfering cells. Full queue was assumed at each transmitting node in the surrounding cells. Thus, the nodes inside the center cell experienced a continuous interference from the interfering nodes in the surrounding cells.

Each cell was sectored into three partitions, the available resources and the number of users is distributed equally on each sector. Each one of the surrounding cells used the same radio resources, i.e., a frequency reuse of one was assumed. There were two examples of relay deployments: two-hop deployment containing six one-hop RS's (Figure 5.4-b) and a three hop deployment with a total of 18 RS's, 6 one-hop RS's and 12 two-hop RS's (Figure 5.4-a). Tables 5.1 and 5.2 contain the ways we performed our grouping as discussed in section 5.7.

We used the same simulation assumptions and parameters used in the previous chapter which are repeated briefly for convenience. We assumed a NLOS link between the BS or an RS and their subordinate UT's and a LOS link between the BS or a RS and another RS if they had a direct father-child relationship, otherwise a NLOS link was assumed. The LOS modified IEEE 802.16 type-D path loss model was used for the BS/RS to RS links with 3.4dB shadowing if they had a fatherhood relationship [14]. The remaining links were assumed NLOS which used the modified IEEE 802.16 path loss model type B with 9.6 dB shadowing standard deviation [14]. SUI 1 model was used for the multipath fading channel connecting the RS's with other RS's and with the BS. For the links to UT's we considered the ITU Ped-A channel model. Refer to Table 3.1 for the other model assumptions.

5.8.2 Results and Discussions

Figure 5.4 presents the throughput performance of our algorithms compared to the reference PF system referred to as CPF. If we examine the performance of the static system, we notice that it is sensitive to the number of users. Indeed, the improvement in the case of 18 users is small compared to the reference system, unlike the case of 72 users, where the improvement is very significant. The adaptive system does not have such a problem and it always shows a very significant throughput improvement, especially at a high traffic load. Considering a full queue model, we can notice up to 65% improvement of 3-hops A-TBSA for the case of 18 UT's, which drops about 27% in the case of S-TBSA. The reason of such drop is the unbalanced

allocation of resources, which causes the system to underutilize resources, hence compromising efficiency. It should be noted that the S-TBSA can be designed from the beginning to match the number of users in each RS. In this case, we should expect a much better performance.

In the case of 72 users, throughput improvement of 3-hops A-TBSA can reach up to 80%, which drops around 20% in S-TBSA. It is can be noticed that TBSA at full queue performs better with a larger number of UT's. This improvement in throughput comes at the price of decreased fairness. Figure 5.5 shows that a significant reduction in fairness at full queue can be experienced in the case of S-TBSA. However, this reduction in fairness is not apparent in realistic load conditions. In fact, we can see our adaptive system provides improved fairness compared to the reference system in most cases. This shows another advantage of adapting resource allocation besides improving the throughput; better throughput fairness can be accomplished.

The statistics of the average user throughput are presented in Figures 5.6 for the two-hop cases, and in Figure 5.7 for the three-hop cases. In Figure 5.6, the CDF of the average user throughput at different load conditions was depicted. At low load conditions, Figure 5.6-a performs almost the same with a slight outage performance reduction. As the load increases, system improvement becomes more apparent. In Figure 5.6-b, A-TBSA was able to cope with the demands of more than 65% of users, which was 2Mbps. The S-TBSA and the CPF can only guarantee to deliver about 1Mps for the same percentage of users. Figure 4.6-C shows the same network demand as 4.6-b, but with a different number of users. Increased number of users improved

both S-TBSA and A-TBSA performance, but to a lesser degree. While the demand which equals 0.5Mbps was delivered to 75% of users for the A-TBSA, the S-TBSA was able to fulfill the demand of 65% of users. This is a significant improvement compared to case 4.6-b for the same network load where the demand by S-TBSA cannot be served for more than 10% of users. Figure 4.6-d shows that all the systems failed to derive the network demand due to system capacity. However, a significant average user throughput was still maintained by our algorithms.

The same conclusions can be drawn from Figure 5.7, which shows the average user throughput for the three-hop cases. In general, better outage and better overall throughput performance was experienced compared to two-hop cases. While there was no difference in performance in Figure 5.7-a, a significant throughput improvement was achieved in the remaining cases. Case b shows the 10^{th} percentile of about 1.8Mbps for A-TBSA, while S-TBSA could not exceed 0.6Mpbs, which is slightly worse than the CPF, which was around 0.7Mbps. In this scenario, the CPF was unable to respond to the demand of any number of users, while S-TBSA guaranteed that 50% of users were served. This improved further in A-TBSA, where the scheduler was able to forward 89% of the user demand. In case (c), S-TBSA improved with increasing the number of users, with 72% of UT's demand being fulfilled in the same way the A-TBSA was able to deliver 95% of its UT's demand. CPF could not deliver the demand of any of its users. Case (d) shows that none of the considered algorithms were able to support such demand. However our solutions were continuously providing better throughput, especially the A-TBSA.

Figures 5.8 and 5.9 present the average throughput performance versus distance for the system with two-hops and three-hops, respectively. Case (a) in both figures shows no significant performance difference between the three algorithms. Case (b) shows a significant improvement for A-TBSA over the two systems. S-TBSA shows some improvement, too, especially at the three-hop case. In both figures 5.8-b and 5.9-b, the average user throughput in the A-TBSA was close to the average demand (2Mbps). The fitting curve lowered in the case of S-TBSA to 1.5Mbps. The CPF suffered from limited resources, which was shown in Figure 5.8-b as a lowered tail close to cell edge, and in 5.9-b as a much lower fitting curve close to 1Mbps, only half of what we achieved via A-TBSA.

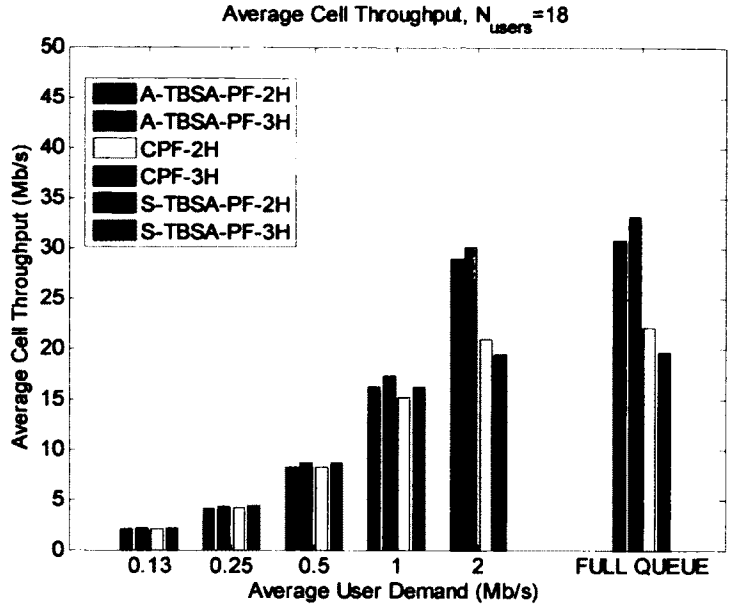
The same can be said about cases (c) and (d), where the improvement of TBSA was obvious due to its IRRR capability, which offered more average throughput to the users throughout the cell. While figures (a), (b), and (c) show an upper limit for the average user throughput, which is located close to the UT's average demand for TBSA and lower than the demand for CPF, case (d) showed a different behavior, where more scattering was apparent in TBSA, a behavior not experienced by CPF. In the cases when the network was unable to deliver all the traffic, the PF scheduler tried to maintain a fair resource allocation. However, our algorithm would scan for any available unutilized resource, even if this may violate the fairness condition, since these resources would be wasted anyway if they remained unused. Because of UT's location randomness, there will be some users who do not experience the same competition on resources compared to the remaining users in the cell. Due to the link

grouping mechanism, the same resources may be assigned to members of the same group, regardless of their traffic conditions. For users who do not experience high competition from their serving node or other access traffic, the scheduler would assign resources to them given that the starving users cannot be served due to capacity limitations on their links. Note that the scheduler will try to balance the allocation to match traffic conditions, but there may still be some users who can enjoy more available resources. This behavior impacts the fairness performance, as shown in Figure 5.5.

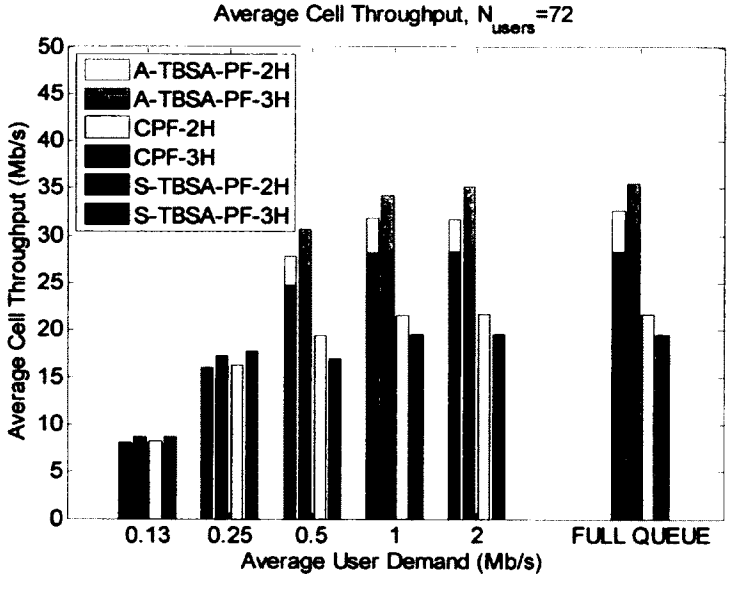
5.9 Conclusions

In this chapter, we have developed a new scheduling algorithm that enables the IRRR. Adopting end-to-end tunneling is an integral part of our algorithm to reduce the complexity of the allocation process; that is, by using tunnel utilization metrics, the flow based allocation process becomes simple while maintaining reuse capabilities. IRRR is supported by link grouping where links can be gathered into groups. The members of a group can be activated simultaneously without overwhelming each other. IRRR improves throughput performance of the TBSA compared to the CPF algorithm.

The pre-allocation algorithm was proposed to adaptively allocate group and zone resources. It provides the original algorithm (S-TBSA) with tuned zone and group allocations. This adaptation process significantly enhances the throughput with a modest increase in algorithm complexity.

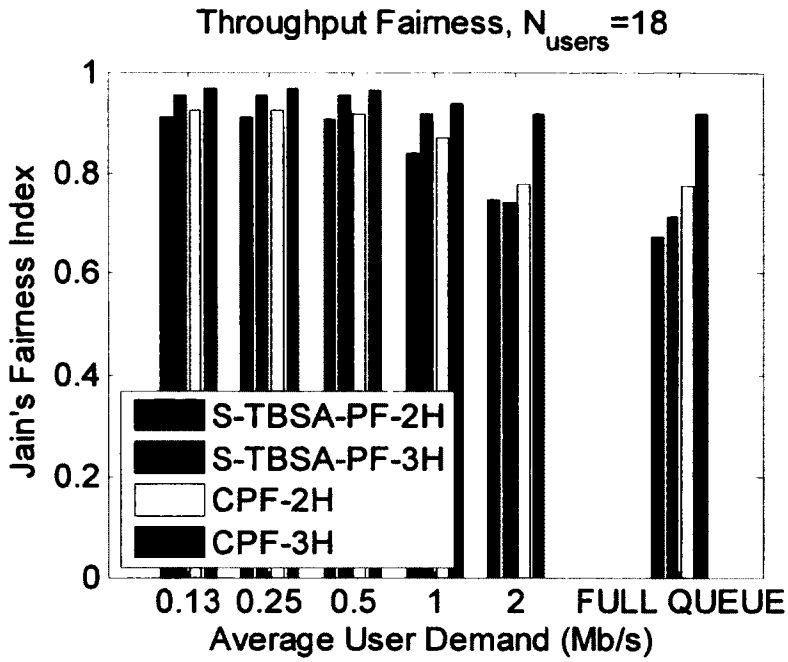


(a)

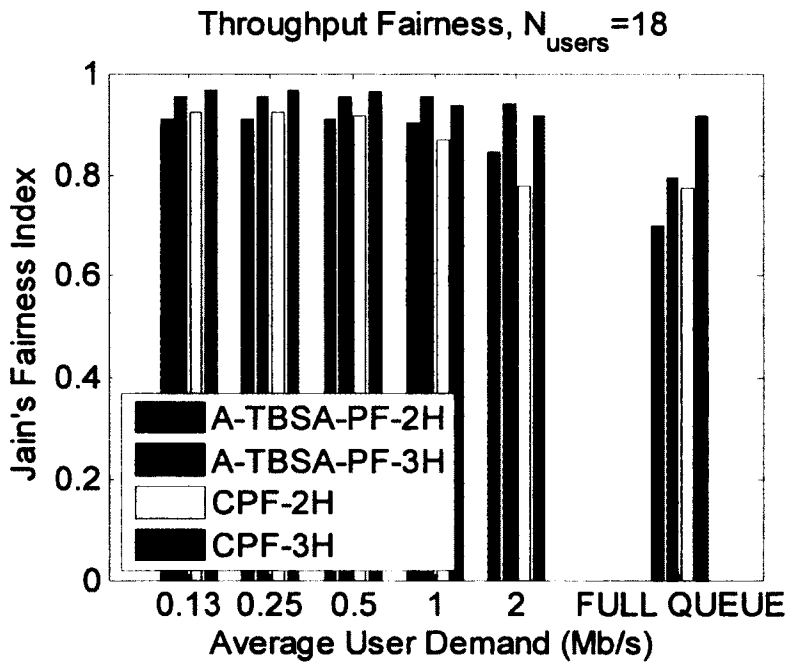


(b)

Figure 5.5 : Average Cell Throughput vs. users demand with different number of users per cell: (a) number of users equal to 18, (b) number of users equal to 72.

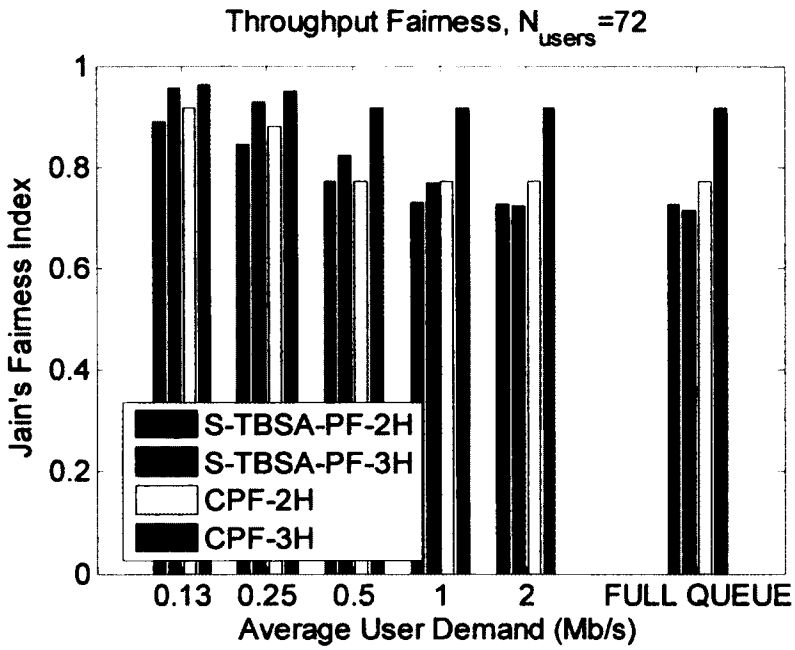


(a)

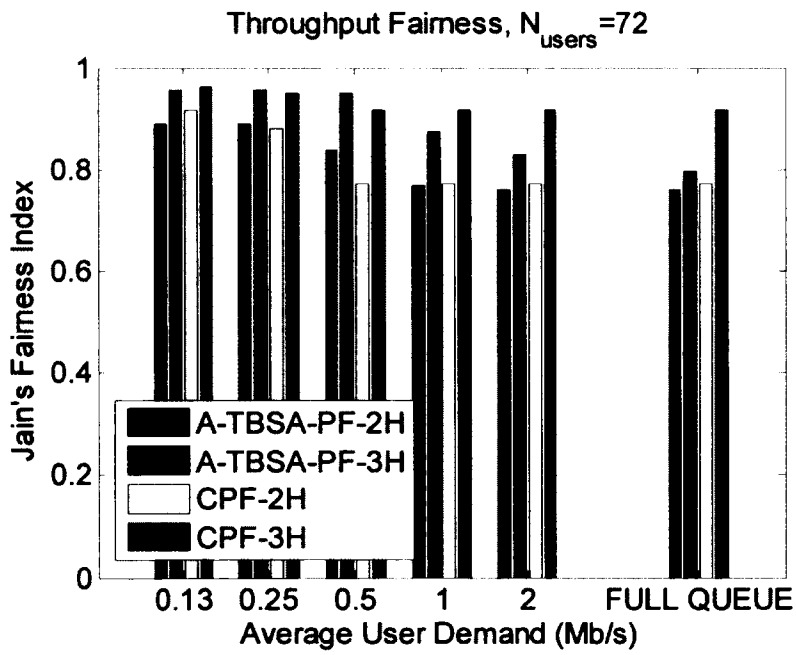


(b)

Figure 5.6 Average Throughput Fairness vs. users demand of the TBSA system with different number of users per cell: (a) S-TBSA with 18 UT's, (b) A-TBSA with 18 UT's (c) S-TBSA with 72 UT's, and (d) A-TBSA with 72 UT's.

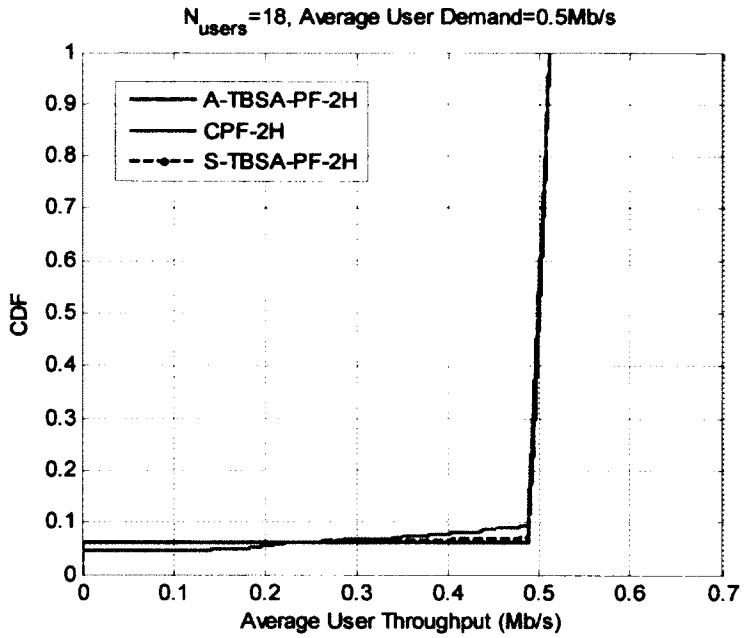


(c)

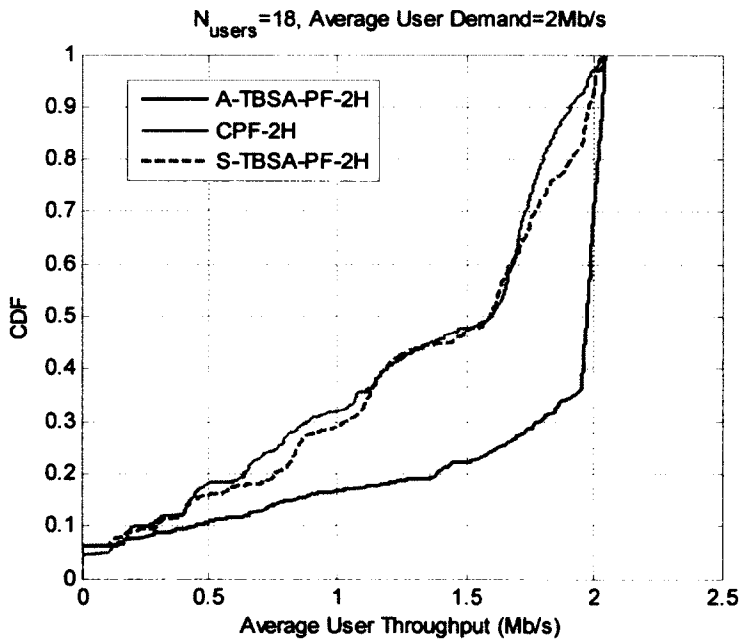


(d)

Figure 5.6 (Cont.) Average Throughput Fairness vs. users demand of the TBSA system with different number of users per cell: (a) S-TBSA with 18 UT's, (b) A-TBSA with 18 UT's (c) S-TBSA with 72 UT's, and (d) A-TBSA with 72 UT's.

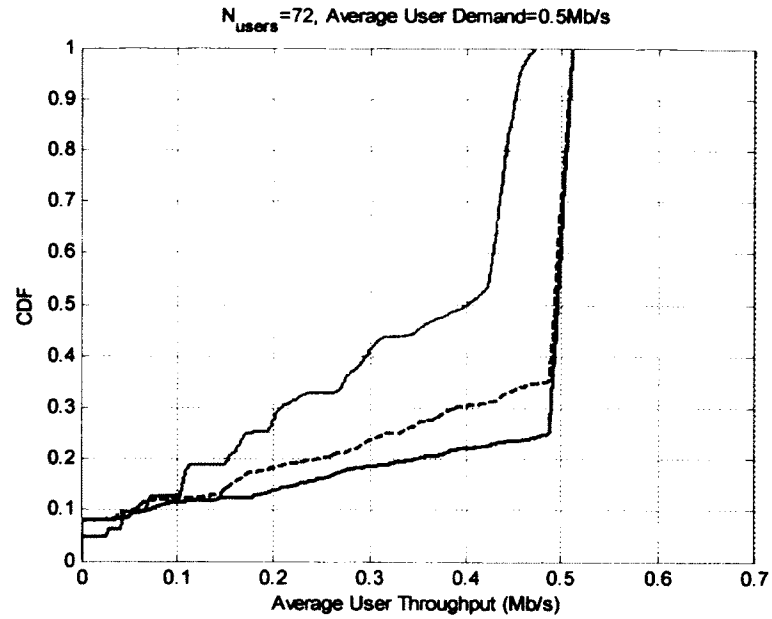


(a)

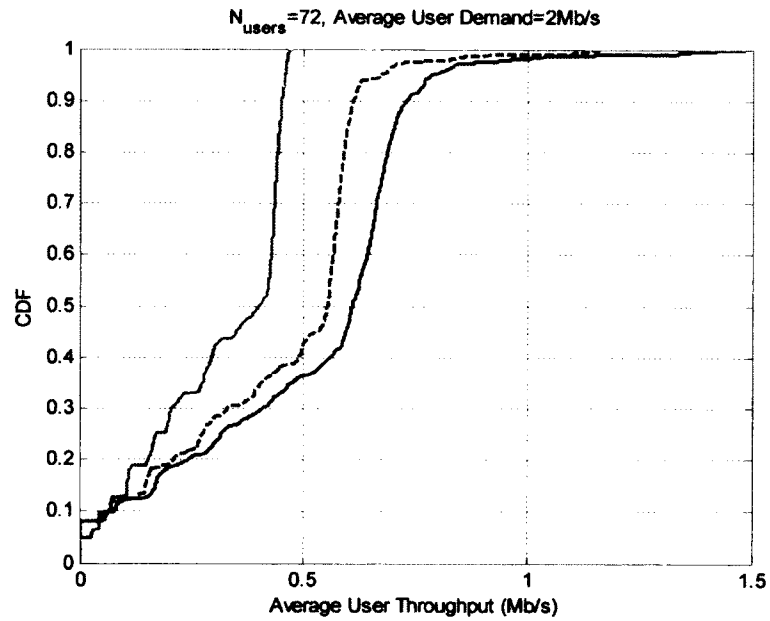


(b)

Figure 5.7 CDF of Average Users throughput of our algorithms considering two-hop cell layout: (a) N_{users}= 18, average demand equals 0.5Mb/s (b) N_{users}= 18, average demand equals 2Mb/s, , (c) N_{users}= 72, average demand equals 0.5Mb/s (d) N_{users}= 72, average demand equals 2Mb/s.

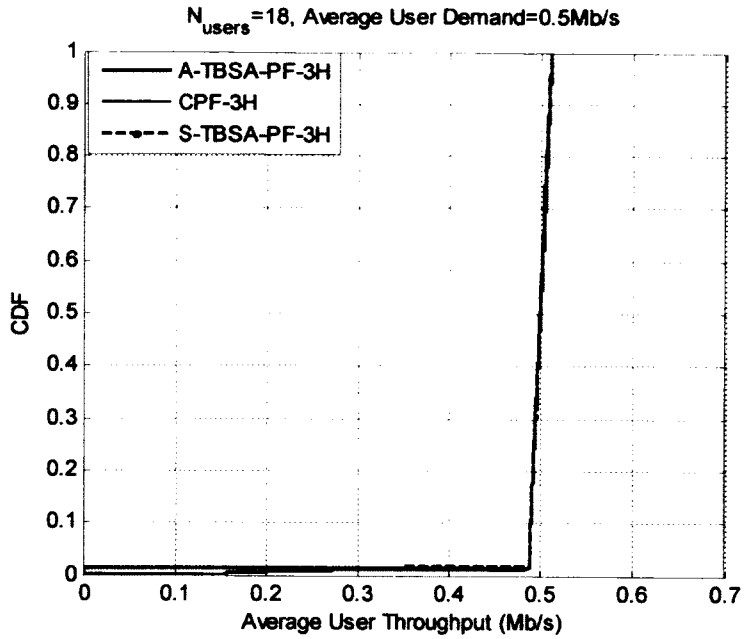


(c)

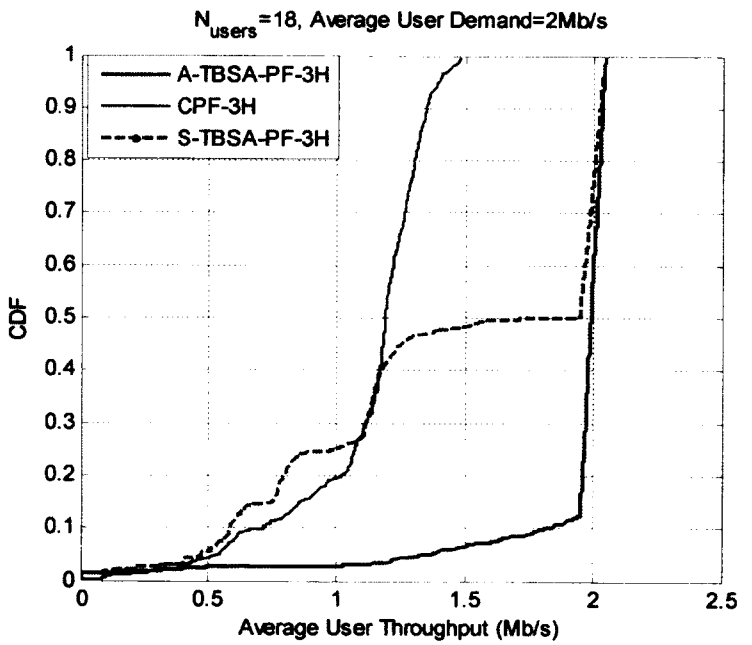


(d)

Figure 5.7 (Cont.) CDF of Average Users throughput of our algorithms considering two-hop cell layout: (a) N_{users}= 18, average demand equals 0.5Mb/s (b) N_{users}= 18, average demand equals 2Mb/s, , (c) N_{users}= 72, average demand equals 0.5Mb/s (d) N_{users}= 72, average demand equals 2Mb/s.

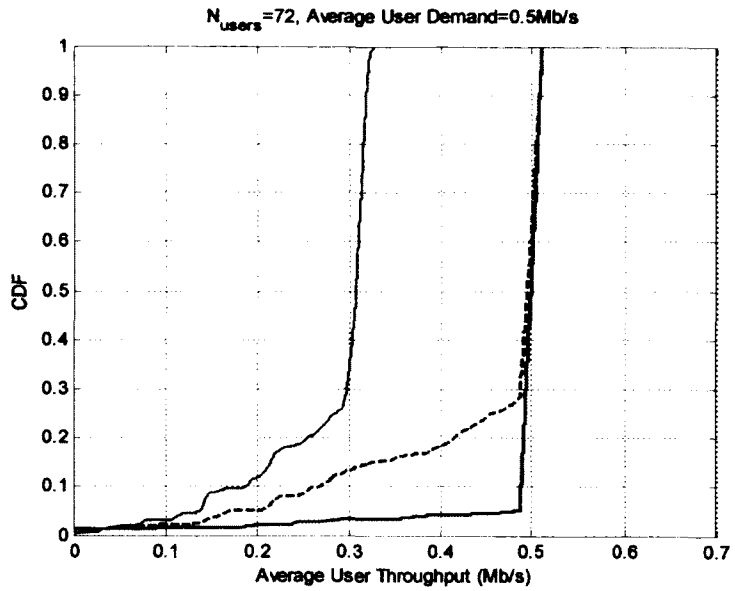


(a)

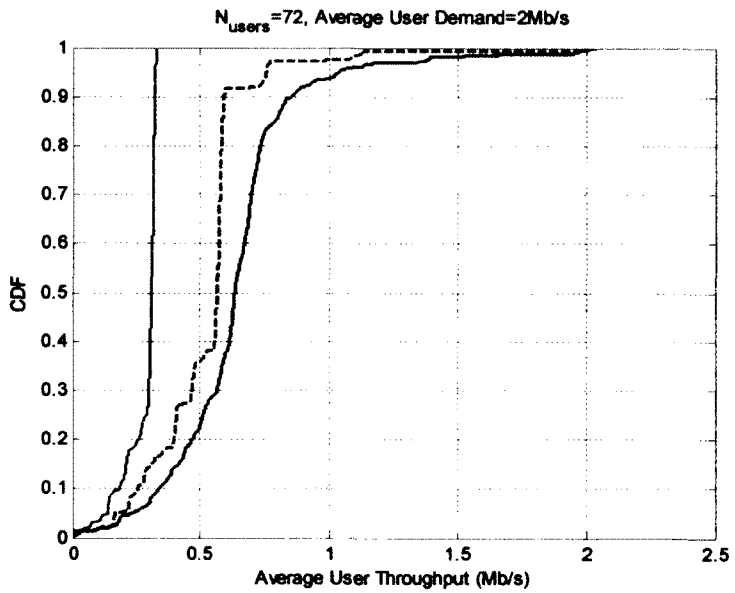


(b)

Figure 5.8 CDF of Average Users throughput of our algorithms considering three-hop cell layout: (a) N_{users}= 18, average demand equals 0.5Mb/s (b) N_{users}= 18, average demand equals 2Mb/s, , (c) N_{users}= 72, average demand equals 0.5Mb/s (d) N_{users}= 72, average demand equals 2Mb/s.

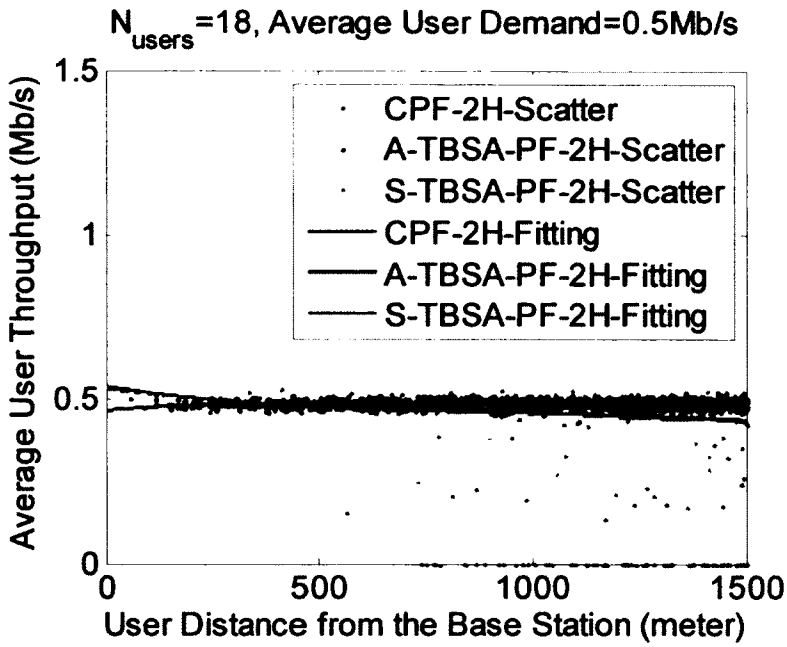


(c)

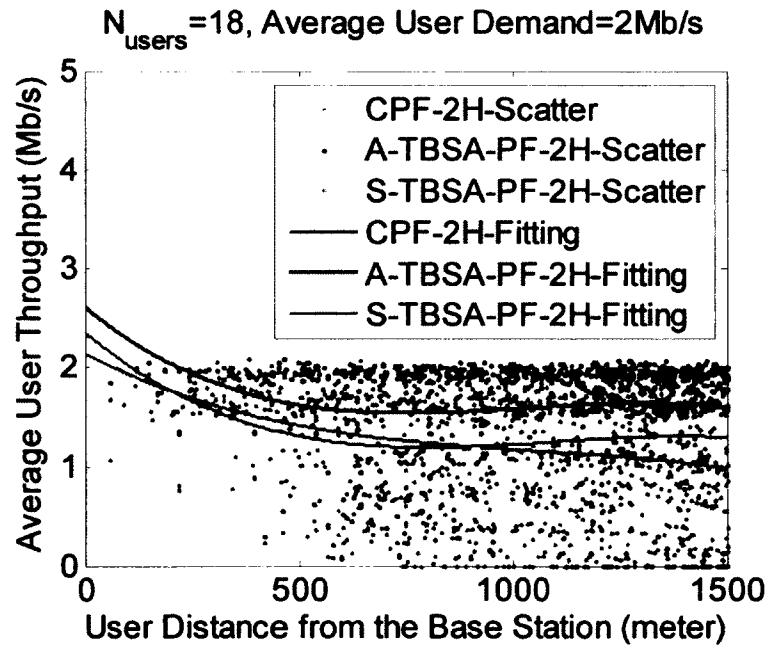


(d)

Figure 5.8 (Cont.) CDF of Average Users throughput of our algorithms considering three-hop cell layout: (a) N_{users}= 18, average demand equals 0.5Mb/s (b) N_{users}= 18, average demand equals 2Mb/s, , (c) N_{users}= 72, average demand equals 0.5Mb/s (d) N_{users}= 72, average demand equals 2Mb/s.

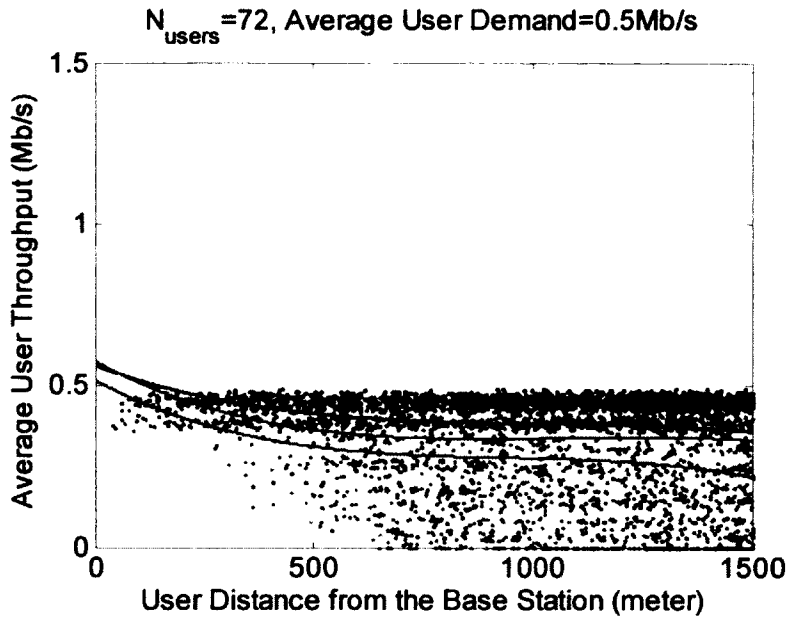


(a)

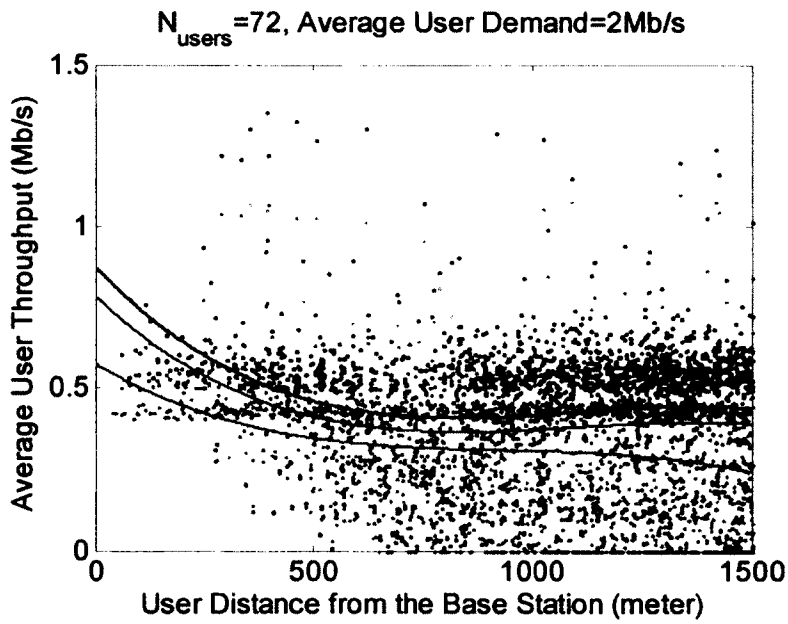


(b)

Figure 5.9 Scatter plot and fitting curves for Average Users throughput vs. distance for cells considering three-hop cell layout: (a) N_{users}= 18, average demand equals 0.5Mb/s (b) N_{users}= 18, average demand equals 2Mb/s, , (c) N_{users}= 72, average demand equals 0.5Mb/s (d) N_{users}= 72, average demand equals 2Mb/s.

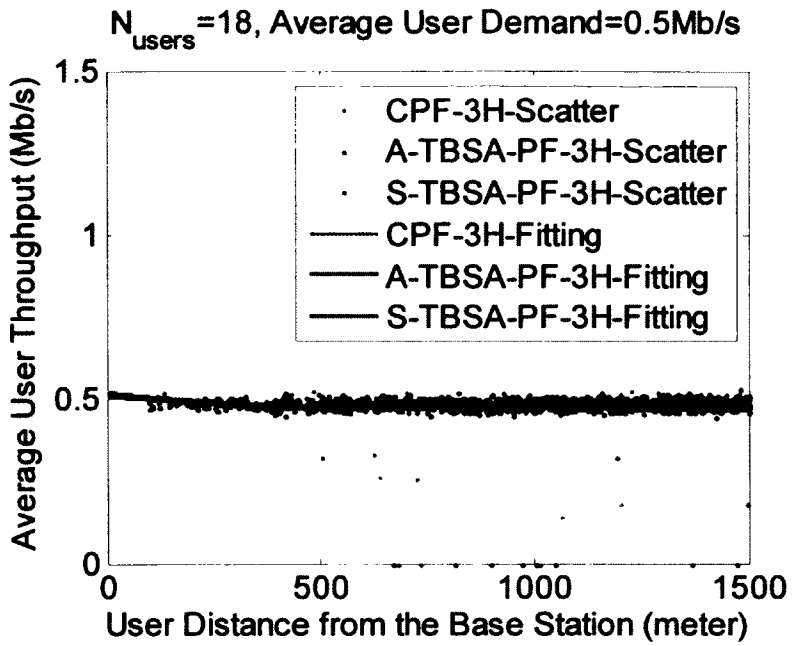


(c)

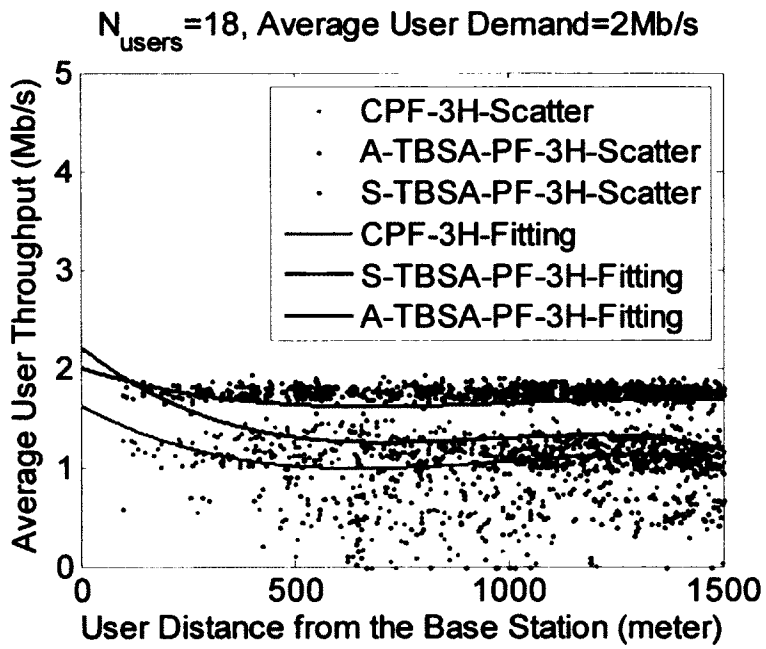


(d)

Figure 5.9 (Cont.) Scatter plot and fitting curves for Average Users throughput vs. distance for cells considering three-hop cell layout: (a) N_{users}= 18, average demand equals 0.5Mb/s (b) N_{users}= 18, average demand equals 2Mb/s, , (c) N_{users}= 72, average demand equals 0.5Mb/s (d) N_{users}= 72, average demand equals 2Mb/s.

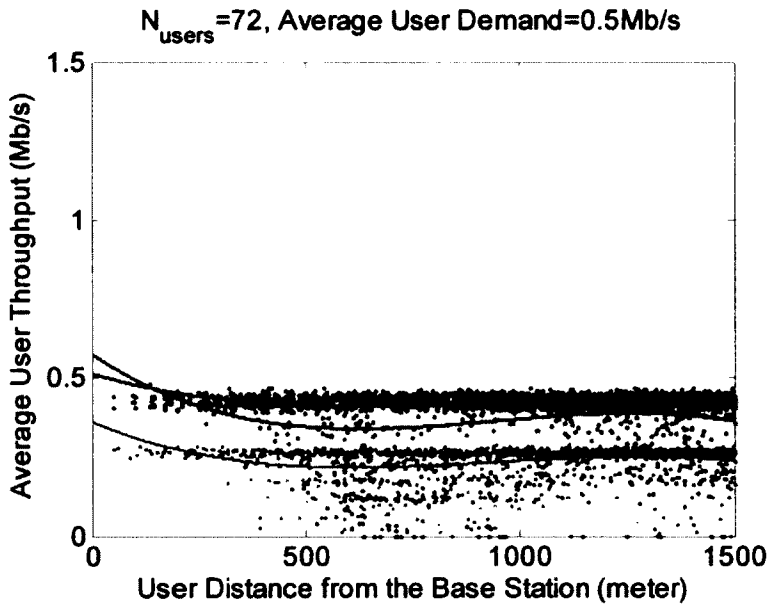


(a)

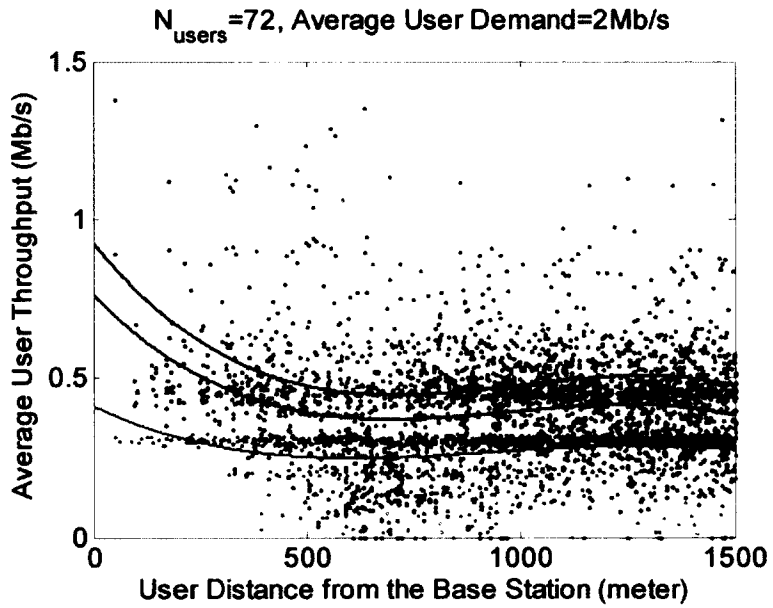


(b)

Figure 5.10 Scatter plot and fitting curves for Average Users throughput vs. distance considering three-hop cell layout: (a) N_{users}= 18, average demand equals 0.5Mb/s (b) N_{users}= 18, average demand equals 2Mb/s, , (c) N_{users}= 72, average demand equals 0.5Mb/s (d) N_{users}= 72, average demand equals 2Mb/s.



(c)



(d)

Figure 5.10 (Cont.) Scatter plot and fitting curves for Average Users throughput vs. distance considering three-hop cell layout: (a) N_{users}= 18, average demand equals 0.5Mb/s (b) N_{users}= 18, average demand equals 2Mb/s, , (c) N_{users}= 72, average demand equals 0.5Mb/s (d) N_{users}= 72, average demand equals 2Mb/s.

Chapter 6

De-centralized Scheduling Algorithm

6.1 Introduction

Centralized scheduling algorithms of multi-hop relay cellular networks suffer from increased overhead as all feedback and other control messages need to be exchanged between the BS and all other nodes in the cell. Centralized algorithms experience large delay in responding to changes in UT access links, as exemplified in the cases when an error event or a steep degradation in link quality occurred. In contrast, de-centralized algorithms can work with much less overhead and can respond quickly to access link changes. Another advantage of de-centralized operation of relays is that they can cognitively scan the spectrum for available bandwidth and utilize it without increased contention on system resources, as we observed in WLMF (Chapter 4) and other centralized algorithms. On the other hand, de-central relaying operation requires a smarter and more expensive RS's.

In recent years, resource allocation for relay based networks has been of significant interest. However, as discussed in the literature, most algorithms suffer

from scalability issues, implying that they are designed for two-hop relay networks only.

A large number of proposals were envisioned as centralized solutions for two-hop relay networks. A greedy distributed solution based on maximal SINR was proposed in [83] to enhance throughput and coverage. An adaptive TDM algorithm was developed to multiplex the traffic of the BS and RS's, and later, FDM was deployed to separate the relay traffic from access traffic. Another greedy solution was proposed in [84]. It assumed two sub-frames; one for BS transmission and the other for RS transmission. The BS performs centralized scheduling for whole traffic. Then, it sends the relayed traffic along with the new schedule. When relays receive the new schedule, they transmit the queued packets first. If the RS receives new relayed traffic, it transmits whenever there are enough resources; otherwise, it is required to ask the BS for more resources.

Another algorithm was implemented by [60] which adopted a priority levels assignment strategy as the scheduling algorithm. Delay, fairness, and channel conditions affected the property level of a queued packet. [85] developed an optimization problem for joint path selection, power, and radio resources allocation. Sub-optimal solutions were developed by continuous relaxation, and an iterative water-filling algorithm.

The solutions presented above were devised for two-hop relay networks and not applicable to a network that features an increased number of relay hops. Some proposals for more hops are also available in extant literature. [77] proposed a three-

hop system where two tiers of relays are placed. The first tier of RS's is fixed while the second tier of RS's is nomadic. An algorithm based on energy saving was implemented.

[9] and [10] proposed centralized based solutions assuming a frame that is partitioned into slots where the relayed traffic could traverse over those slots within a frame. Clearly, this is not realistic and, as we showed in chapter 4, the multi-frame was the remedy to such a problem.

As in chapter 5, IRRR is supported by using relay grouping. According to our knowledge, there is no scalable scheduling algorithm devised for flows with more than two-hops that can support intra-cell resources reuse in OFDMA relay based cellular networks. Here, we have to distinguish our system from other multi-hop networks such as mesh networks or sensor networks. The algorithms built for such networks assume no centralized entity which can globally allocate the resources in the cell. Furthermore, solutions built for single cell networks, such as [12], are mainly routing based solutions where finding the path is integrated with a scheduling process. Such an approach increases the overhead, complicates the allocation process, and introduces increased delay.

A de-centralized multi-hop solution that assumes full queue traffic was proposed in [60] that allows the RS to schedule its arriving packet before it arrives. The RS initially calculates the PF metric for its UT and decides the flow to be scheduled and the size of traffic it can accept. It then transmits this allocation schedule to its father node, which in turn checks the available resources and decides if it can process an

acceptance. If yes, it relays this information to its father and the process goes on until the new schedule arrives to the BS. The main flaw of such a proposal, similar to that in [9] and [10], is that the transmission from the BS to the last relay hop is assumed be done within the duration of a frame. This assumption is not realizable because a minimum of one frame is required to relay traffic from one hop to the other. Another problem is that [60] assumes static resources allocation to RS's, regardless of traffic or link quality conditions, which reduces resources utilization. Moreover, [60] did not provide answers to critical queries, such as:

- In cases when the pre-scheduled flow does not have a queued packet. How should the BS respond in such a situation?
- If the quality of the access or relay link changes in a manner that violates the pre-scheduled allocation, how should the RS react?
- If a queue builds up at an RS, what should the RS do?

In this work, we propose de-centralized algorithms for OFDMA relay based cellular networks. These algorithms maintain queue stability by instructing the relays to report their queue status and allocate resources accordingly. They are scalable as they maintain a linear complexity. More importantly, they are able to enhance utilization by means of IRRR via link grouping.

6.2 Contributions

1. We propose a distributed algorithm titled "Distributed Scheduling based on Queue Status" or (DSQS). The following are the main features of the DSQS algorithm:

    a. This solution requires a minimum interaction from the BS. The BS is required to respond to the RS's request for additional bandwidth. Minimum processing and overhead requirements are needed at the BS.

    b. Resource reuse is supported to improve resource utilization.

2. We propose a novel distributed algorithm named Distributed Scheduling based on Generalized Fair Sharing (DSGFS), which defines a merit for each relay to be used for fair resource allocation to all relays. The main features of DSGFS are:

    a. High utilization can be achieved by engaging the BS in the scheduling process.

    b. It is possible to enhance throughput performance with much less overhead and processing requirements compared to centralized schemes.

    c. Fair resource sharing based on merits, where the resources are allocated based on this merit relative to the merits of other relays in the cell.

    d. Intra-cell resources reuse enabled by means of link grouping to enhance throughput.

6.3 Distributed Scheduling based on Queue Status (DSQS).

We now consider a scenario where the BS is not required to schedule the flows individually; rather, it will aggregate the flows according to the links that connect the BS to its direct RS's. In this scheme, the direct RS's are the customers. These RS's may be owned privately and can be equipped with the capability to use free spectrum. When no free spectrum is available, the RS sends a message to its father node asking for more resources. The father node – if it is not the BS – will forward this message back to the BS, which is required to allocate resources to each RS.

The RS's allocate the resources to its associated access and relay links in a distributed fashion. They are required to forward the traffic to both child nodes and access UT's. For fair comparison with other solutions, we will consider a case where no free spectrum is available, although in practice the RS's may find free spectrum, which will ultimately enhance the throughput. In the following sections, we will describe our system, explain our algorithm, and provide a simple enhancement to its relaying delay performance.

6.3.1 System Description

We assume an OFDMA cellular network containing multiple cells. Each cell has a single BS located at the center and is surrounded by multiple relays. Each RS receives relay traffic from its father node, which could be the BS or an RS. The routes from the BS to the RS's are well established before any scheduling process takes place. The UT's select the serving node based on the received SINR. However, this is not a

requirement in our algorithm. In fact, a more sophisticated relay selection algorithm can be adopted such as [73] or [74] where end-to-end throughput can be maximized. As mentioned earlier, we will adopt SINR as our relay selection method because of its simplicity and ability to perform well if relays are well positioned [74].

The BS assigns a queue for each direct RS (not all RS's). All the traffic going through that direct RS will enter that queue whether the destination UT is served directly by that RS, or by one of its children RS's. The BS or any RS will treat the aggregated traffic as if it is a flow of a normal user. The scheduler at any node can be viewed as a single-hop scheduler.

As in all relay systems, dedicated channels or TDM partitions are required for transmitting all relay control messages from/to the BS and all father RS's to their subordinate RS's. No other special frame modification is required by our algorithm. Thus, we can assume that two dimensional radio resources units (slots) are available at our cell. The BS is required to allocate resources for its access traffic and all relay traffic. Furthermore, the BS will assign radio resources to RS's when those RS's have backlogged traffic.

Our algorithm supports one QoS class only, namely "best effort". QoS can be supported via classifying the traffic according to QoS requirements. Separate problems for each class can be implemented. Classes are ranked according to their priority where real-time classes have top priority and best effort traffic has the lowest. The scheduler first formulates the problem for the highest ranked traffic. Then, the next

highest ranked class is served and so on. In our work, we will consider a single class and down link traffic only.

6.4 Scheduling based on Queue Status

The DSQS algorithm is based on RS queue status. The BS queues are not considered in our algorithm as they are mainly dependent on incoming traffic conditions. Unlike BS queues status, RS queue status is mainly affected by scheduling and resource allocation performed at the BS. The RS reports the status of its queue to the BS, indicating the number of slots, $\mathbf{s}_y(m)$, required to empty the queued access and relay traffic which can be calculated as

$$\mathbf{s}_y(m) = \left\lceil \sum_{l \in L(y)} \frac{Q_l(m)}{q_l(m)} \right\rceil \tag{6.1}$$

where $L(y)$ contains all the links served by node y, i.e. $L(y) = \mathcal{U}(y) \cup \mathcal{D}(y)$. $Q_l(m)$ is the queue size in bits of link l, which can be an access link or a relay link and $q_l(m)$ is the link spectral efficiency in bits per slot. To avoid wastage of resources, the RS should exclude the backlogged traffic of UT's if they are in outage. Although not supported in our implementation, it would be beneficial if the RS could instruct the BS to stop forwarding the traffic of the outage UT's.

When the BS receives a message from RS y asking for resources, it allocates resources for the needy RS first before scheduling new traffic. New allocations for RS's need to be synchronized. Thus, the influence of the received allocation message

has to be delayed, such that all relays can receive the new assignment. For example, if RS y with access links of hop count equal to h_y requests $\mathbf{s}_y(m)$ slots, where m is the time at which RS y sends its request, the BS will receive this request at $m + h_y - 1$. Allocation for this request has to be delayed $h_{max} - h_y$ frames to synchronize group allocation changes at all hops. If τ is the time at which new allocations take place, i.e. τ is the application time, the time of the allocation at the BS will be v,which is given by

$$v = \tau - h_{max} + 1 \tag{6.2}$$

We are assuming the BS is not involved in flow scheduling and knows neither the link conditions, nor traffic conditions. The BS will only forward the aggregated traffic to the next link. Each node, whether it is a BS or a RS, knows its access users and next link children RS. Father nodes are required to determine the next nodes to which they will send the relay traffic and perform the required scheduling. If a RS cannot transmit some or the entire queued packets, it is required to send a message back to the father relay asking for extra resources.

The status of queues is assumed to be reported to the BS at every frame when there are queued packets that cannot be transmitted due to an insufficient number of available resources. However, reporting queue status at every frame would increase the overhead and hurt the efficiency of the system. It is possible to reduce this overhead by decreasing the frequency of queue reporting, but that may reduce the throughput and increase the delay and queue size.

Since it is assumed that the BS will not track the traffic and channel conditions of each UT's flow, it cannot optimize flow allocation procedure as in centralized approaches. Therefore, channel unaware allocation procedure following resource fair strategy is adopted to fairly distribute the resources.

At scheduling instance v, the BS will first assign the requested slots to each RS $\mathbf{s}_y(m)$ before any new flow scheduling at the BS. The new assignment will take place atτ. By allocating the RS's the requested resources before any new traffic scheduling, we guarantee the stability of RS queues. We can say that a queue is stable when its size is bounded, and this stability can be met if the scheduler assigns the RS the requested number of slots to empty its queue before scheduling new traffic. Therefore, we can be certain that the queue will not grow indefinitely. This is the basic idea of the DSQS algorithm.

6.4.1 Intra-cell Radio Resources Reuse (IRRR)

To enhance the utilization and improve the throughput performance, we support intra-cell radio resources reuse (IRRR) by grouping the nodes according to their ability to transmit simultaneously, such that the UT's or child RS's served by nodes in a group will not be overwhelmed by the interference from other nodes in the same group. This grouping is assumed fixed during the experiment. A single node, multiple nodes, or even a sector of a node can be a member of a group. We will refer to a member of a group as Transmit Entity (TE) indicating that the grouping is based on transmitters

that can operate simultaneously without causing interruption to the operation of each other.

The BS will add the resource requirement of each node (or a sector of a node) to compute the resource requirement of the TE, which may comprises multiple AN's will require the number of resources $\mathbf{s}_a(m) = \sum_{y \in a} \left(\mathbf{s}_y(m)\right)$. The group assignment denoted by $\mathfrak{S}_g(\tau)$ is equal to the maximum required number of resources of its members, which is given by

$$\mathfrak{S}_g(\tau) = \max_{\forall a \in g}\left(\mathbf{s}_a(m)\right) \tag{6.3}$$

where $m = \tau - h_{max} - h_a + 1$.. Note that $\mathfrak{S}_g(\tau)$ cannot exceed the number of available resources. The number of available resources should be updated to make sure the allocated resources do not exceed the available resources, as shown in Algorithm 6.1 where we present our DSQS in more detail.

By responding to RS queue requirements, we can avoid overflow of relay queues by assigning enough resources for queued packets. Furthermore, new traffic scheduling can be reduced by decreasing the allocated resources of new traffic to fulfill the increased demand of RS's on resources to empty their queues. If the total allocation capacity is not sufficient to accommodate the waiting packets, the RS's with higher hop count are assigned the requested resources first. It should be noted that relay grouping may create queue instability especially if an RS and its serving BS sector belong to the same group. Any resources given to that RS to empty the queue will be allocated to its serving BS sector which may result in a highly unstable

condition where the resources are locked for that sector. Therefore, additional steps are needed to avoid such a situation. One method would be to avoid having any child/father relationship in the group. This condition has already been satisfied for two and three hops. The second option is to cease allocating additional relay traffic by the grandfather BS sector.

If there are remaining resources, i.e. $\mathfrak{S}(\tau)>0$, the BS will utilize these resources to enhance utilization. The BS can increase the number of resources to send more access traffic and relay traffic. Due to grouping mechanisms, the BS will be treated as multiple nodes, i.e. each BS sector constitutes an AN. One or more sectors can join a group and each one will be assigned resources as any other AN in the same group (see Figure 6.1).

Fair allocation strategy is implemented to divide the resources between access and relay traffic. Defining j as an RS that is directly connected to the BS and designating N_j as the total number of users served by RS j or by one of its children RS's, the average throughput of a BS link is computed according to

$$\bar{r}_l(v) = \left(1 - \frac{1}{t_c}\right)\bar{r}_l(v-1) + \frac{1}{t_c} r_l(v-1) \tag{6.4}$$

where l could be a relay link $j \in \mathcal{D}(BS)$ or an access link $i \in \mathcal{U}(BS)$, and $r_l(\tau)$ is the throughput previously scheduled for link l during $\tau - 1$, which is equal to the number of allocated bits for each direct relay link j.

Our procedure calculates a link fairness factor, $\alpha_l(\tau)$, which is equal to $\alpha_{j^*}(\tau) = N_{j^*}/\bar{r}_{j^*}(\tau)$ for relay links and $\alpha_{A_{0,S}}(\tau) = \sum_{i \in U(0,S)}\left(1/\bar{r}_i(\tau)\right)$ for the access traffic at BS

sector $(0,S)$, where N_{j^*} is the number of all UT's served directly or indirectly by j^*. The overall fairness factor of the sector is the sum of fairness factors of its links $\alpha_{0,S}(\tau) = \sum_{l=L(0,S)} \alpha_l(\tau)$. If two or more BS sectors are members of the same group, the maximum fairness factor of these sectors will constitute the group fairness factor, which is calculated as $\alpha_{g^*}(\tau) = \max_{(0,S)\in g^*}\left(\sum_{l=L(0,S)} \alpha_l(\tau)\right)$. The group allocation is normalized according to $\dot{\alpha}_{g^*}(\tau) = \alpha_{g^*}(\tau)/\sum_{\forall g} \alpha_g(\tau)$, and multiplied by the total number of resources to get the fair allocation of group resources. However, as was stated earlier, RS queue requirements have to be fulfilled before the group allocation takes place. Therefore, allocation of a group is equal to at least $\mathfrak{S}_g(\tau)$ (Eq. 6.3). If there are remaining resources, we may increase the number of allocated resources such that either the fair allocation, $S_g(\tau)$, has been granted or there are no more available resources.

The fair allocation, $S_g(\tau)$, is proposed to be equal to the normalized fairness factor $\dot{\alpha}_{g^*}(\tau)$ multiplied by the total number of resources, S, i.e. $S_g(\tau) = \dot{\alpha}_{g^*}(\tau)S$. The procedure sorts the groups starting from the one with the highest fairness factor, and allocate these with a maximum of two allocations: 1) the allocation based on queue requirement, $\mathfrak{S}_g(\tau)$, and 2) the allocation based on resource fairness, $S_g(\tau)$; $\mathfrak{S}_g(\tau)$ has already been granted. The extra resources may be assigned for fairness-based allocation subject to resource availability as shown in Algorithm 6.1-a.

TE's that are members of the same group will be scheduled the same group resources whereas AN's within a TE will be allocated a fraction of TE allocation

according to their resource requirements. For example, the allocation of AN y belonging to TE a will be calculated as

$$\mathfrak{S}_y(\tau) = \frac{\mathbf{s}_y(m)}{\mathbf{s}_a(m)}\mathfrak{S}_a(\tau) \tag{6.5}$$

For the flows that are served directly by the BS, the allocation will be based on fairness requirement. Thus, the allocated resources for the access traffic of BS sector $(0, S)$ can be written as

$$\mathfrak{S}_{A_{0,S}}(\tau) = \frac{\alpha_{A_{0,S}}(\tau)}{\alpha_{A_{0,S}}(\tau) + \sum_{j\in\mathcal{D}(0,S)}\alpha_j(\tau)}\mathfrak{S}_{g(a)}(\tau) \tag{6.6}$$

In the same manner, we can denote the allocation of a relay link as

$$\mathfrak{S}_{j^*}(\tau) = \frac{\alpha_j(\tau)}{\alpha_{A_{0,S}}(\tau) + \sum_{j\in\mathcal{D}(0,S)}\alpha_j(\tau)}\mathfrak{S}_{g(a)}(\tau) \tag{6.7}$$

It is important to emphasize that the allocation that has been carried out at time v is for future application (at time $\tau = v + h_{max} - 1$). Flow scheduling at the BS is based on previously allocated resources $\mathfrak{S}(v)$ on frame time $v - h_{max} + 1$. Flow-based utility scheduling procedure can be used to forward the access traffic at BS sector (0, S) such that the number of resources do not exceed $\mathfrak{S}_{A_{0,S}}(v)$ while the FIFO principle is used for relay traffic. See Algorithm 6.1 for more details.

6.4.2 Allocation at the AN's

Before allocating traffic to their access UT's, the RS's are required to forward all relay traffic based on the FIFO principle (Algorithm 6.1-c). This is required to reduce the

delay of the higher hop count UT's. The remaining resources will be utilized for access traffic.

Now the allocation problem is partitioned into smaller problems at each relay. We define $\boldsymbol{X}$ as a matrix of size $\mathcal{Z} \times \mathfrak{S}$ containing the variables $\{x_{\mathfrak{z}s}(v)\}$. The optimization problem for each AN can be written as

$$\max_{\boldsymbol{X}} U_a(v)$$

Subject to

$$\begin{gathered} \sum_{\forall \mathfrak{z} \in \mathcal{U}(a) \cap \mathrm{D}(a) \cup a} \sum_{\forall s \in \mathfrak{S}(a)} x_{\mathfrak{z}s}(v) \leq \mathfrak{S}_a(v) \\ \sum_{\forall \mathfrak{z} \in \mathcal{U}(a) \cap \mathrm{D}(a) \cup a} x_{\mathfrak{z}s}(v) \leq 1 \\ x_{\mathfrak{z}s}(v) \in \{0,1\}, \mathfrak{z} \in \mathcal{Z}, s \in \mathfrak{S} \end{gathered} \tag{6.8}$$

$U_a(v)$ can be written as

$$\begin{aligned} U_a(v) &= \sum_{i \in \mathcal{U}(a)} \sum_{\forall s \in \mathfrak{S}(a)} x_{is}(v)\, U_{is}(v) \\ &= \sum_{\forall s \in \mathfrak{S}(a)} U_s(v) \end{aligned} \tag{6.9}$$

where $U_s(v) = \sum_{i \in \mathcal{U}(a)} x_{is}(v)\, U_{is}(v)$. This problem is a typical generalized assignment problem which is known to be NP-hard [87]. An approximation algorithm, or search-based solution, can be used to solve this allocation problem. In our work, we rely on a greedy approximation algorithm by assigning the slots to the UT with the highest utility. Clearly, each slot can now be allocated independently from each other, further simplifying the allocation by maximizing the utility of each slot alone.

Therefore, we can write the allocation problem per slot as

$$\max_{\{x_{i,s}(v)\}} U_{is}(v)$$

$$\text{Subject to} \quad \sum_{\forall s \in \mathfrak{S}(a)} x_{is}(v) \leq 1 \tag{6.10}$$

And the solution will be

$$x_{is}^{*}(v) = \begin{cases} 1 \; if \; i = \arg\max_{i} U_{is}(v) \\ 0 \; if \; i \neq \arg\max_{i} U_{is}(v) \end{cases} \tag{6.11}$$

Implementing slot-by-slot allocation is inefficient. Hence, when a flow is selected for transmission, we propose to let the scheduler schedule a number of contiguous slots to the same flow; otherwise the addressing overhead will be very large. Thus, the solution will be modified such that a flow is selected based on an average utility. The selected flow may be allocated the available resources until all of its queued packets are transmitted, or until there are no more available resources. In this case, the solution would be as follows:

A flow is selected for scheduling according to:

$$i^{*} = \text{argmax}_{i}\big(U_i(k)\big) \tag{6.12}$$

The data size to be transmitted will be:

$$D_{i^{*}}(k) = \text{minp}\big(W_{i^{*}}, q_{i^{*}}\mathfrak{S}_{a(i^{*})}\big) \tag{6.13}$$

And accordingly the number of slots can be calculated as:

$$\hat{x}_{i^{*}}(k) = \left\lceil \frac{D_{i^{*}}(k)}{q_{i^{*}}} \right\rceil \tag{6.14}$$

The allocation goes on for the next flows until there are no more backlogged users or free resources. Refer to Algorithm 6.1-d for more details.

It is worth noting that utility maximization within a RS does not have a significant performance influence. Figure 6.1 shows unnoticeable performance difference among the three utility functions: proportional fairness (PF), maximum throughput (MT), and maximum throughput fairness (MF) applied at each RS while the BS allocation procedure is kept the same in all cases. The reason is the increased availability of access resources due to our relay grouping mechanism. It is understood that if sufficient resources are available for scheduling, any utility function will result in the same throughput performance. System assumptions for Figure 6.1 are presented in section 6.5. In the next section, we will present our grouping strategy and why we have increased resource availability at the higher hop count RS's.

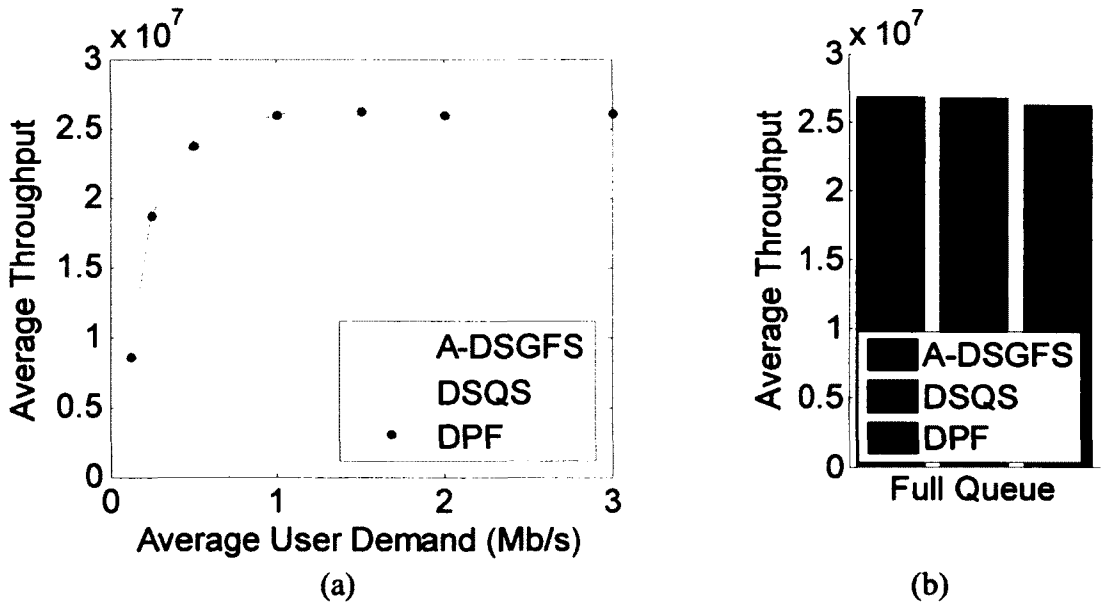


Figure 6.1 Average Cell Throughput of different utility assumption at RS scheduler for three-hop relaying case with 72 UT's with different traffic conditions: (a) Poisson arrival with different average users demand, (b) Full Queue.

Algorithm 6.1 DSQS Algorithm

(a) Resource Allocation At the BS during v

The following is the resources allocation that will be activated at time $\tau = v + h_{max} - 1$

$s_a(v - h_a) = \sum_{y \in TE(y)} s_y(v - h_y)$; On v the BS has received the queue status message sent by RS y during frame time $(v - h_y)$. The required resources by TE a is the sum of all resources required by nodes to which they belong.

$\mathfrak{S}(\tau) = S$; All resources are available at the beginning.

For all a starting from the higher hop count

$\delta_a = \min\left(\max\left(0, s_a(v - h_a) - \mathfrak{S}_g(\tau)\right), \mathfrak{S}(\tau)\right)$;The TE assignment should not exceed the available resources

$\mathfrak{S}_{g(a)}(\tau) = \max\left(\delta_a + \mathfrak{S}_{g(a)}(\tau), \mathfrak{S}_{g(a)}(\tau)\right)$; Group allocation based on RS queue requirement

$\mathfrak{S}(\tau) = \mathfrak{S}(\tau) - \delta_a$;Update the number of remaining resources

END If

$\alpha_{j^*}(\tau) = \frac{N_{j^*}}{\bar{r}_{j^*}(\tau)}$; Fairness factor for relay links.

$\alpha_{A_{0,S}}(\tau) = \sum_{i \in U(0,S)} \frac{1}{\bar{r}_i(\tau)}$; Fairness factor for BS sector $(0, S)$ access traffic.

$\alpha_{g^*}(\tau) = \max_{(0,S) \in g^*}\left(\sum_{l=L(0,S)} \alpha_l(\tau)\right)$; We sum all factors that belong to the same sector. If two or more sectors are members of the same group, select the maximum.

$\alpha_{g^*}(\tau) = \frac{\alpha_{g^*}(\tau)}{\sum_{\forall g} \alpha_g(\tau)}$

$S_{g^*}(\tau) = \alpha_{g^*}(\tau) S$; Group Allocation based on Fairness factors

g_sorted = SORT($\{\mathfrak{S}_g\}$,'Ascend') ;

$Count = N_g$;N_g is the number of groups

While$\mathfrak{S}(\tau)$>0 & Count>0

$g = g\_sorted(Count)$;Select the group with the highest resources requirements

Count= Count-1;

$\Delta = \min\left(S_g(\tau), \mathfrak{S}(\tau) + \mathfrak{S}_g(\tau)\right)$; % Fair based allocation $S_g(\tau)$should NOT exceed the remaining resources.

$\mathfrak{S}_g(\tau) = \max(\mathfrak{S}_g(\tau), \Delta)$; % Group allocation may increase to enhance fairness but must not decrease to fulfill RS queue requirement.

$\mathfrak{S}(\tau) = \min(\mathfrak{S}(\tau), \mathfrak{S}(\tau) + \mathfrak{S}_g(\tau) - \Delta)$; Update the number of remaining resources

End While

$\mathfrak{S}_a(\tau) = \mathfrak{S}_{g(a)}(\tau)$; The resources available for each TE.

$\mathfrak{S}_y(\tau) = \frac{s_y(m)}{s_a(m)} \mathfrak{S}_a(\tau), y \in TE\ a$; The resources available to each TE are divided over its nodes based on their reported resource requirement.

$\mathfrak{S}_{A_{0,S}}(\tau) = \frac{\alpha_{A_{0,S}}(\tau)}{\alpha_{A_{0,S}}(\tau) + \sum_{j \in D(0,S)} \alpha_j(\tau)} \mathfrak{S}_{g(a)}(\tau)$;The resources for access traffic is scaled based on fairness factor

$\mathfrak{S}_{j^*}(\tau) = \frac{\alpha_j(\tau)}{\alpha_{A_{0,S}}(\tau) + \sum_{j \in D(0,S)} \alpha_j(\tau)} \mathfrak{S}_{g(a)}(\tau)$The resources for relay traffic is scaled based on fairness factor

(b) Relay traffic at the BS

For all relay link $j^* \in L(y)$ DO

$FIFO(W_{j^*}(v), \mathfrak{S}_{j^*}(v))$; Packets are selected based on FIFO basis from queue $W_{j^*}(v)$ such that the allocated slots does NOT exceed $\mathfrak{S}_{j^*}(v)$.

END DO

(c) Relay Traffic at each RS y

For all relay link $f^* \in L(y)$ DO
$FIFO(W_{f^*}(v), \mathfrak{S}_y(v))$; Packets are selected based on FIFO basis from queue $W_{f^*}(v)$ such that the allocated slots does NOT exceed $\mathfrak{S}_y(v)$.
Calculate $\mathfrak{S}_y(v)$; Calculate the number of resources available for access traffic
END DO

(d) Flow based access traffic allocation at node y [1]

Given $\mathfrak{S}_y(v)$
$Count = N_t$
While $Count > 0$ & $\mathfrak{S}_y > 0$
$Count = Count - 1$

$$t^* = \underset{t}{\text{argmax}}\, U_t(k), \text{ where } t \in U(y)$$

$$D_{t^*}(k) = \text{minp}\left(W_{t^*}, q_{t^*}\mathfrak{S}_{y(t^*)}\right)$$

$$\mathfrak{X}_t(k) = \left\lceil \frac{D_{t^*}(k)}{q_{t^*}} \right\rceil$$

END While

[1] ***The same procedure used at each BS sector***

6.4.3 AE Grouping in DSQS

Unlike the TBSA, the DSQS does not assume zone-based frame structure for traffic allocation. The complete 2-D resource units are available for both relay and access traffic. The AN can schedule the relay traffic similar to a UT link. Addressing and controlling messages require dedicated channels or zones. A common assumption in our algorithm is that at least two zones, or two dedicated channels are required for the nodes to transmit and receive management information. The case is different for data traffic, where a zone-based allocation dedicates a zone for all relay traffic, whereas the data traffic in DSQS is scheduled freely in the available 2-D resource units. A silent period equal to one time symbol is assumed before and after allocating traffic to subordinate RS to allow them to switch from transmit to receive status and vice versa.

In DSQS, each RS is allocated a number of resources for both relay as well as access traffic. Distributed scheduling is performed by each AN for both relay traffic and access traffic by utilizing the resources that have been allocated by the BS. IRRR is supported in DSQS by means of Transmitter Entities (TE) grouping. The grouping is for the TE's that can *transmit* simultaneously without overwhelming their receivers. In comparison, the grouping in the TBSA is based on the links that can be activated simultaneously, i.e. the grouping is based on the entities that can receive simultaneously. A member of a group could be a single node, a sector of a node, or a number of RS's. When a number of RS's form an entry in a group, they share resources (non-reusable sharing).

Typically, the relay links are expected to tolerate the interference better than the access links due to the assumed superior RS equipment, as well as their anticipated optimized placement compared to UT's. Thus, a higher intra-cell reuse gain is expected in the relay zone compared to the access zone. On the other hand, zoning has a negative impact on resources availability for the higher hop count links where relay zones are partially used or even unused. This is addressed in the DSQS system as all the resources allocated to a group are completely assigned for every member in that group. Therefore, the higher hop count *access* links have a larger number of available resources, which results in better utilization.

The same grouping used for access links in the TBSA will be used for the DSQS. We provide references to two examples presented in chapter 5. They are repeated here for the sake of convenience. Figure 6.3 shows an example of a single-group allocation. If a zone-based frame structure is implemented as in Figure 6.3-b, the relays at higher hop counts will suffer from low utilization. This is unlike case 6.3-a, where more resources are available for access traffic at a higher hop count.

The relays are grouped based on the interference and half-duplex constraints. Assuming TE j is a node that serves the direct RS's designated in the set, $\mathfrak{D}(j)$, and the access UT's, $\mathcal{U}(j)$. $\mathcal{U}(j)$, we can define the set $L(i)$, which contains all the links served by node j, i.e. $L(j) = \mathcal{U}(j) U \mathfrak{D}(j)$. .

The following constraints should be maintained in relay grouping:

1. Interference Constraint

The receivers of any group member should be able to receive from their father node without a significant interference from other members of the same group. If $TE\ a \in g_T$, then $TE\ j \in g_T$ if $SIR_{L(j)\leftarrow a} > \widetilde{g_T}\, SIR_{min}, \forall a, j \in g_T$, which implies that the signal to interference ratio between the transmissions received by two RE's in a group is lower bounded by $\widetilde{g_T}\, SIR_{min}$. This is given by

$$SIR_{L(j)\leftarrow a} = \min\left(\frac{P(L(j)\leftarrow j)\mathfrak{h}(L(j)\leftarrow j)}{P(L(j)\leftarrow a)\mathfrak{h}(L(j)\leftarrow a)}\right) \tag{6.15}$$

where $P(L(j)\leftarrow j)$ is the transmitted power of TE j and $\mathfrak{h}(L(j)\leftarrow j)$ is the channel gain from the TE j to $L(j)$. This bound is increased as the number of elements in the group $\widetilde{g_T}$, increase to account for the increased interference.

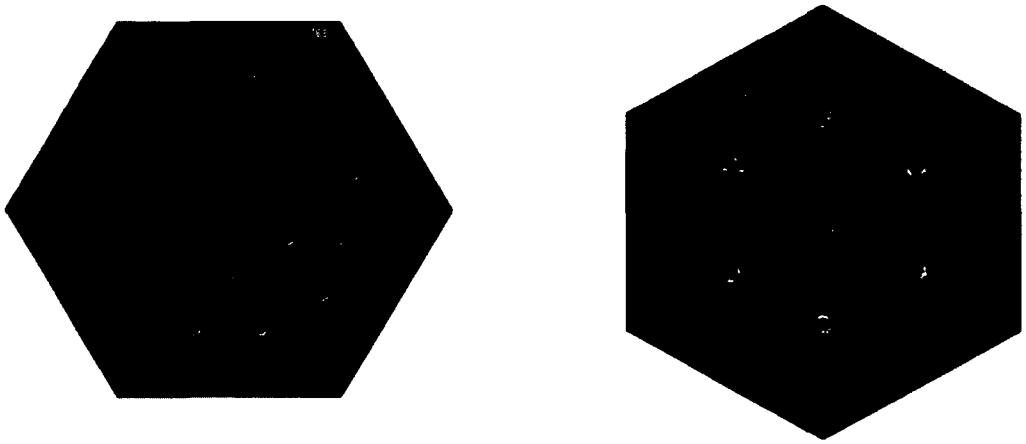
2. Half-Duplex constraint

A Half-duplex operation is assumed for each node, i.e. simultaneous transmission and reception is not allowed. Therefore, the direct father of the TE a, $\mathfrak{F}(a)$, and the direct child $\mathfrak{D}(a)$, are not allowed to join the group due to half-duplex constraint. That is, if $TE\ a \in g_T$, then $TE\ j \in g_T$ if $j \neq \mathfrak{F}(a) \& j \neq \mathfrak{D}(a), \forall j, a \in g_R$.

The DSQS works with any grouping satisfying the two conditions stated earlier. To evaluate performance, we have examined two examples of cell layout and partitioning scheme depicted in Figure 6.2 and Table 6.1. The cell is sliced into a

number of partitions where each partition joins a group. The partitions of the same group carry the same color. The partition may include a RS, a sector of the BS or a group of RS's. Table 6.1 presents the list of partitions and their TE index number, as well as the group to which each TE belongs.

The allocated resources of groups can be based on TDM or FDM. TDM may be considered a better choice in this scenario. This is because when a TE is in receiving mode, all of its FDM allocated resources will be wasted. FDM, however, can enhance the flexibility of assigning resources due to its high number of sub-channels. Thus, we adopt a TDM/FDM frame partition scheme (Figure 6.4-c) where FDM between pairs in groups do not have father/child relationships among their members. Thus, we avoid the aforementioned wastage. Conversely, we employ TDM between pairs containing members that have child/father relationships.



(a) (b)

Figure 6.2 Cell portioning schemes: (a) three-hop case, and (b) two-hop case.

Partition Index	TE Index
(1,1)	1
(2,1)	2
(3,1)	3
(4,1)	4
(5,1)	5
(6,1)	6
(1,2)	7
(2,2)	8
(3,2)	9
(4,2)	10
(5,2)	11
(6,2)	12
(1,3)	13
(2,3)	14
(3,3)	15
(4,3)	16
(5,3)	17
(6,3)	18

(a)

Partition Index	TE Index
(1,1)	1
(2,1)	2
(3,1)	3
(4,1)	4
(1,2)	5
(2,2)	6
(3,2)	7
(4,2)	8
(1,3)	9
(2,3)	10
(3,3)	11
(4,3)	12

(b)

TE Groups	Member TE		
1	1	9	5
2	2	12	16
3	3	11	13
4	4	8	18
5	5	7	15
6	6	10	14

(c)

TE Groups	Member TE	
1	1	7
2	2	4
3	3	9
4	4	6
5	5	11
6	8	10

(d)

Table 6-1 Link grouping the two cases shown in Figure 6.2: (a) Tunnels and RE indices for three-hop partitions, (b) Tunnels and RE indices for three-hop partitions and (c) Relay groups and their members for three-hop case, and (d) Tunnels and RE indices for three-hop partitions.

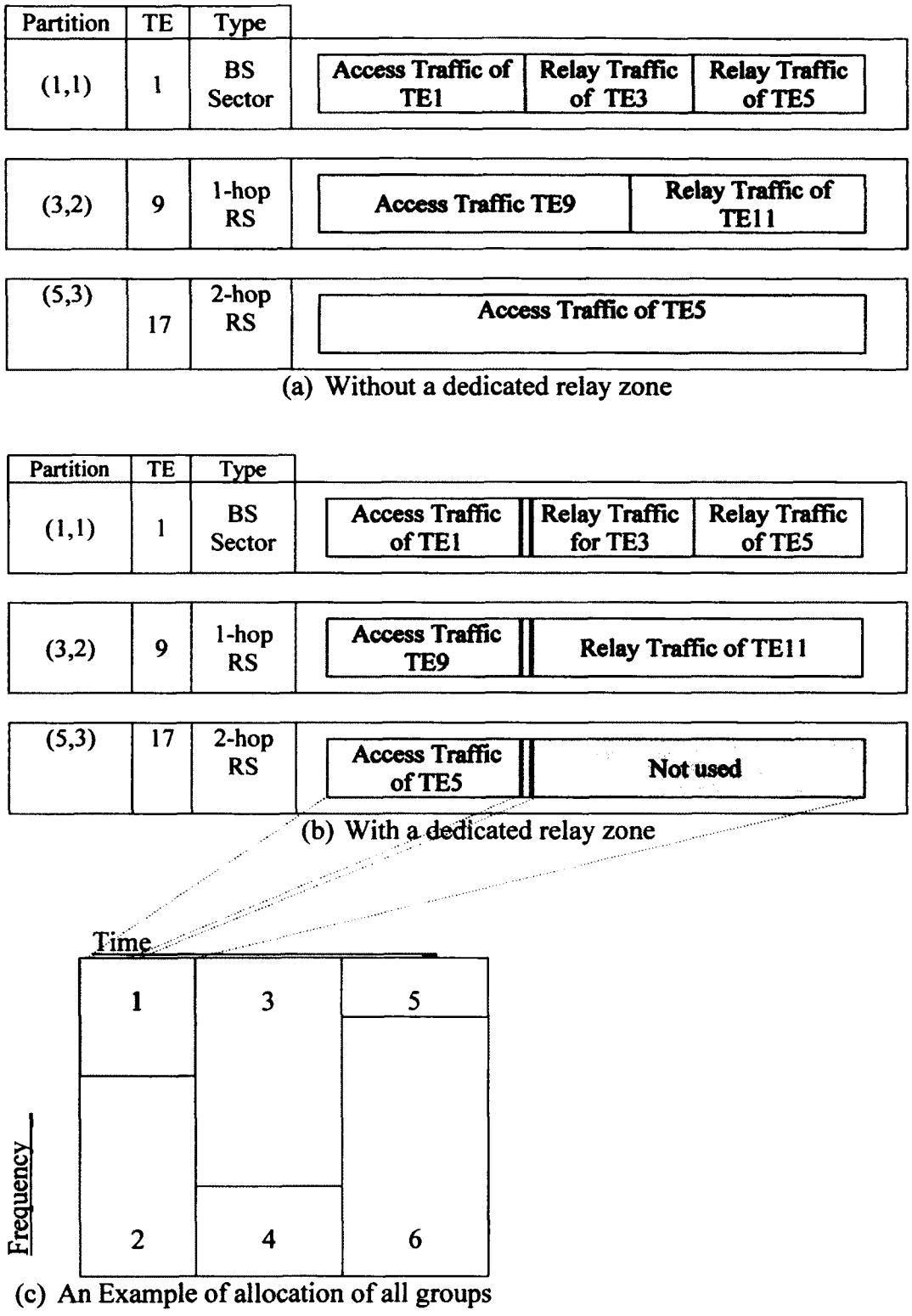


Figure 6.3 An example of group allocation (a) with no dedicated zone, (b) with a dedicated relay zone.

6.5 Distributed Flow Scheduling Based on Generalized Fair Sharing

Distributed allocation has its benefits in reducing overhead and providing fast response to UT link changes. The utilization in the DSQS algorithm previously introduced was not optimized because we did not provide any mechanism to ensure the effective reuse of every possible resource unit. Furthermore, matching resource allocation to the forwarded traffic will enhance resource utilization and improve throughput.

In this section, we will introduce a new de-centralized algorithm that, with insignificant feedback overhead, will be able to schedule flows from the BS based on the decision of the scheduler at the RS's. Intra-cell reuse is supported using link grouping, as was observed in chapter 5, to enhance the utilization and improve throughput performance. In the following section, we will describe the system and explain our algorithm.

6.5.1 System description

In this scheme, we assume that the BS queues the arrived packets according to their access RS. We adopt *the zone-based frame structure*, where two zones are assumed at both downlink and uplink sub-frames. The BS is required to allocate an appropriate number of resources for both zones.

We define an Access Entity (AE) which can be an AN, a group of AN's, or may be a sector of an AN. An end-to-end tunnel is assumed between the BS to every AE

over which the BS aggregates all traffic of the access node of that AE. The AN's that belong to an AE are responsible for allocating the resources to their access UT's. As per our scheduling strategy, the AN's will expect the BS to deliver packets along with sufficient resource allocation to accommodate those packets.

Resources for access links and relay links must be allocated before accepting new traffic to ensure that the scheduler is capable of delivering all packets without encountering a large relay queuing delay. We first allow the scheduler to decide on the allocation of all links, followed by forwarded traffic based on this allocation process. The relays are informed by the new allocation prior to the arrival of new packets. This is done in a manner that the relay is able to schedule the arrived packets according to the received allocation message. The BS assigns two allocations to each RS: one for access traffic and the other for relay traffic. In the following sections, we describe these two procedures.

6.5.2 Resource allocation of access zone

In the access zone, each AN requires a number of radio resources, enough to forward the access traffic to their UT's. Two features have been proposed for the purpose of allocating resources for access traffic: *generalized fairness concept* and *link grouping*.

6.5.2.1 Generalized Fairness Concept

Fairness among nodes can have different interpretations. Throughput fairness and cost fairness between AE's or UT's are examples of such fairness interpretations. Generally speaking, we define a local *merit* as a measure used to evaluate fairness at each AE, which will be used as the basis on which resources will be divided between those AE's.

Different fairness interpretations can be supported. Table 6.1 lists some merits and their objectives. In our work, we propose such a generalized fairness concept which depends on the local merit designed to enhance a certain fairness performance. Among the merits listed in Table 6.1, proportional fair (PF) merit can achieve a better compromise between throughput and fairness. Therefore, we have adopted proportional fair merit in our work.

6.5.2.2 Flow list

Since the BS is not involved in flow scheduling, we expect the scheduler to perform in a round robin manner. A round robin scheduler is not a good performer in terms of throughput, as it is channel quality unaware. Typically, proportional fair strategy achieves a better compromise between throughput and fairness[1]. Therefore, we propose to create a flow list for each RS to list the best candidate flows. We allow the AN to inform the BS with the best candidate flow based on a local optimization procedure. The BS keeps this flow on top of the list, which is continuously updated. The BS scans this list to schedule a flow. The one at the top of the list is selected first,

followed by the second, and so on, until either no more packets are queued, or the allocated resources are fully consumed. If a flow is served and all of its packets have been fully emptied from the queue, it is moved to the bottom of the list. This is to allow the other flows an opportunity to be selected, which will in turn enhance the fairness.

As we have used the user PF merit presented in Table 6.1, we will also use the flow Proportional Fairness (PF) index given by $q_i/\bar{r}_i$ to rank the flow. The flow with the highest rank is given a higher scheduling priority. Other rankings can be adopted based on the desired objective. For example, Maximum throughput Fairness (MF) index $1/\bar{r}_i$, can be used for flow ranking when throughput fairness achieves merit. On the other hand, merits based on cost fairness are better used with the PF index.

6.5.2.3 Intra-cell Radio Resource Reuse (IRRR)

To enable IRRR, the AE's will be grouped into what we refer to as access groups. Nodes in a group can transmit simultaneously to their UT's, such that the UT's will not be overwhelmed by the interference from the AE's in the same group. This grouping is assumed fixed and is determined by system design. Assuming that the AE's are arranged into $\mathfrak{N}$ groups, each group will get a portion of the available radio resources in the access zone. All nodes within a group will get the same access radio resources.

For each group g_A we will define the *group merit* as m_{g_A}, which is equal to the maximum local merit among all AE's in the same group g_A to maintain fairness and

load balancing, i.e.,$m_{g_A} = \max_{a \in g}(\varrho_a)$. The resources in the access zone will be divided over those groups according to their merits. If the total number of slots in the access zone is denoted by $\mathfrak{S}$, then group $g_A$will be allocated $\mathfrak{S}_{g_A} = \alpha_{g_A}\mathfrak{S}$ slots, where $\mathfrak{S}$ is the total available resource for local access traffic and α_{g_A} is the portion of the resource assigned to group g_A determined according to group merits

$$\alpha_{g_A} = \frac{m_{g_A}}{\sum_{\forall g_A} m_{g_A}} \tag{6.16}$$

The BS is the entity responsible for AE's resource assignment. Referring to relaying with decentralized scheduling, each AN will be responsible for scheduling its UT's in the assigned radio resources. Scheduling of the UT's by the AN will be done in the same manner as done by the BS in a single-hop case.

Merit	**Objective**
1	Site cost fairness
$1/\bar{r}_z$	Site throughput fairness
N_z	User cost fairness
$\sum q_i/\bar{r}_i$	User proportional fairness
$N/\bar{r}_z$	User throughput fairness

Table 6-2 Different Merits and their objectives

6.5.3 Scheduling in the Relay Zone

As we stated earlier, each access node a will be assigned its group resources $\mathfrak{S}_{g_A(a)}$ if $a \in g_A(a)$. The number of bits transmitted by AN a cannot exceed $\mathfrak{B}_a = \mathfrak{S}_{g_A(a)}\bar{q}_{A_a}$, where $\bar{q}_{A_a}$ is an estimation of average access link quality, $\bar{q}_{A_a} = \frac{1}{\grave{\mathcal{U}}(a)}\sum_{i\in\mathcal{U}(a)}\bar{q}_i$, where $\mathcal{U}(a)$, $\grave{\mathcal{U}}(a)$ and $\bar{q}_i$ are the set, the number, and access link quality of UT's served directly by node a, respectively.

In the relay zone, radio resources are allocated for the backhaul traffic, i.e., RS/BS to RS traffic. In our work, we assume half-duplex BS and RS's. Therefore, a RS cannot transmit and receive simultaneously. To support IRRR, relay link grouping is adopted as discussed in the previous chapter. A group is formed by combining the links that can be activated simultaneously with tolerable interference. We will distribute the resources according to the requirement of each node access traffic, $\mathfrak{B}_a$, specified earlier. To transmit $\mathfrak{B}_a$ bit from the father RS $\mathcal{F}(a)$ to node a, we need $\lceil \mathfrak{B}_a/q_a \rceil$ slots, where q_a is the average spectral efficiency of the link between $\mathcal{F}(a)$ and its child AN a. The required number of slots is calculated for all relay links in the same group and the maximum slots needed are assigned as the group resource requirement.

Theorem 6.1

If we assign α_{g_A} as the portion of resources assigned to access group g_A, and assume that the zone size can be adaptively changed, we may write the zone size as $\mathbb{Z}_A = \frac{1}{\gamma+1}\mathcal{S}$, for access zone, and $\mathbb{Z}_R = \frac{\gamma}{\gamma+1}\mathcal{S}$ for a relay zone, where $\mathcal{S}$ is the total number of resources and γ, a variable that depends on link qualities and local merits, and is given by

$$\gamma = \sum_{\forall g_R} \max_{\forall a \in g_R} \left(\frac{\alpha_a q_{A_a}}{q_a} + \sum_{\forall h \in \mathfrak{R}(AS_a)} \frac{\alpha_h q_{A_j}}{q_a} \right) \tag{6.17}$$

Proof

The proof of this theorem will follow after propositions 6.1, 6.2, and 6.3.

Proposition 6.1

If we denote the set of all relays served by node a or any of its subordinate nodes by $\mathfrak{R}(a)$, node a will require a number of slots to *receive* the relayed traffic from its father AN. This is illustrated as follows:

$$\mathbf{s}_a = \left\lceil \frac{\mathfrak{B}_a}{q_a} + \sum_{\forall j \in \mathfrak{R}(a)} \frac{\mathfrak{B}_j}{q_a} \right\rceil \tag{6.18}$$

Note that $\mathfrak{B}_j = \mathfrak{S}_j q_{A_j}$, where $\mathfrak{S}_j$ is the radio resource of the node j.

Proposition 6.2

Since $\mathfrak{S}_j = \alpha_j \mathfrak{S}$, we can rewrite $\mathbf{s}_a$ as $\mathfrak{S}\,\beta_a$, where β_a is independent of the available slots rather, it is related only to the merits and link quality. It can be shown as

$$\beta_a = \frac{\alpha_a q_{A_a}}{q_a} + \sum_{\forall h \in \Re(AS_a)} \frac{\alpha_h q_{A_j}}{q_a} \tag{6.19}$$

Proposition 6.3

The relay zone size can be calculated as $\mathbf{Z}_R = \mathfrak{S}\gamma$, where γ is given by Eq. 6.19.

Proof

For each relay group g_R, we calculate the number of required slots denoted by $\mathfrak{S}_{g_R}$as the maximum number of resources required by its member nodes

$$\begin{aligned} \mathbf{s}_{g_R} &= \max_{\forall a \in g_R} (\mathbf{s}_a) \\ &= \mathfrak{S} \max_{\forall a \in g_R} (\beta_a) \\ &= \mathfrak{S}\,\beta_{g_R} \end{aligned} \tag{6.20}$$

We then add the number of slots required for all groups to identify the size of the relay zone $\mathbf{Z}$. This can be expressed as

$$\begin{aligned} \mathbf{Z}_R &= \sum_{\forall g_R} \mathbf{s}_{g_R} \\ &= \mathfrak{S} \sum_{\forall g_R} \beta_{g_R} \\ &= \mathfrak{S}\gamma \end{aligned} \tag{6.21}$$

γ is a variable that is independent to the available resources; it is dependent only on channel qualities and relay merits. This is shown as follows:

$$\gamma = \sum_{\forall g_R} \beta_{g_R} \tag{6.22}$$

Q.E.D.

Proof of Theorem 6.1

The allocation of resources should not exceed the available slots, i.e.

$$\mathbb{Z}_R + \mathbb{Z}_A \leq \mathcal{S} \tag{6.23}$$

where $\mathbb{Z}_A$ equals $\mathfrak{S}$ whereas $\mathbb{Z}_R$ equals $\mathfrak{S}$ multiplied by γ. Eq. 6.23 can be rewritten as:

$$\mathfrak{S}\gamma + \mathfrak{S} \leq \mathcal{S} \tag{6.24}$$

Therefore, by scaling $\mathfrak{S}$, we can get the highest possible resource utilization. The highest possible access region size $\mathfrak{S}$ can be written as

$$\mathfrak{S} = \frac{\mathcal{S}}{\gamma+1} \tag{6.25}$$

The resource assignment of each zone is equal to

$$\mathbb{Z}_A = \mathfrak{S} = \gamma_A \mathcal{S} \text{ for Access Zone,} \quad \text{where } \gamma_A = \frac{1}{\gamma + 1} \tag{6.26}$$

$$\mathbb{Z}_R = \gamma\mathfrak{S} = \gamma_R \mathcal{S} \text{ for Relay Zone,} \quad \text{where } \gamma_R = \frac{\gamma}{\gamma + 1} \tag{6.27}$$

Q.E.D.

Theorem 6.1 enables us to find the number of resources for each zone. Dynamically, the zone allocation can change based on the changes in merits we defined earlier. This is not expected at every frame; rather, we prefer the changes to be made every number of frames due to overhead associated with the changes of slot assignments.

6.5.4 AE Resource allocation

After the number of slots is found for each zone, we identify the number of slots at each node. The number of slots allocated to each node at the relay zone can be written as

$$\mathbf{s}_a = \mathbf{s}_{g_R(a)} = \beta_{g_R(a)}\mathbf{Z}_R/\gamma \tag{6.28}$$

Access traffic at each node belonging to group g_A will be assigned a resource number equal to:

$$\mathfrak{S}_a = \mathfrak{S}_{g_A(a)} = \alpha_{g_A(a)}\mathbf{Z}_A \tag{6.29}$$

The amount of data traffic to be forwarded for each AN will be limited by $\alpha_{g_A}\mathbf{Z}_A$ as shown in section 6.5.7.

6.5.5 AN Resource allocation

The AE may contain one or more AN's. The resources of AN y can be calculated by appropriate scaling of AE resources. Thus, we can express the allocated resources as

$$\mathbf{s}_y = \frac{\beta_y}{\beta_a}\mathbf{s}_a \text{for allocation in Relay Zone} \tag{6.30}$$

$$\mathfrak{S}_y = \frac{\varrho_y}{\varrho_a}\mathfrak{S}_a \text{ for allocation in Access Zone} \tag{6.31}$$

6.5.6 Static Zone Allocation

Assuming that the relay zone is static, the above procedure would still be applicable. However, at the last step, we would need to keep the same ratio between the allocations of each zone. We denote the fixed access and relay zones assignments by $\dot{\mathbf{Z}}_A$, and $\dot{\mathbf{Z}}_R$, respectively. These zone assignments can be written as $\dot{\mathbf{Z}}_A = \dot{\gamma}_A \mathcal{S}$ and $\dot{\mathbf{Z}}_R = \dot{\gamma}_R \mathcal{S}$, where $\dot{\gamma}_A$ and $\dot{\gamma}_R$ are the portion of resources for the access and relay zones respectively. We identify the bottleneck zone to determine the maximum possible traffic that can be scheduled by the BS. The bottleneck zone can be calculated by dividing over γ_A, and γ_R, respectively. The one that results in the minimum ratio is the bottleneck which would determine the allocation of other zones.

Proposition 6.4

The bottleneck zone is the zone that has a lower allocation compared to the adaptive version identified by Theorem 6.1:

$$Z_{min} = \underset{z}{\operatorname{argmin}} \left(\frac{\dot{\gamma}_Z}{\gamma_Z} \right) \tag{6.32}$$

where $Z \in \{A, R\}$.

Proposition 6.5

The static zone allocation can be written as

$$\mathbf{Z}_z = \dot{\gamma}_{Z_{min}} \frac{\gamma_Z}{\gamma_{Z_{min}}} \mathcal{S} \tag{6.33}$$

Thus, we can express the static allocations for the access zone as:

$$\mathbf{Z}_{\mathrm{A}} = \begin{cases} \dot{\gamma}_A \mathcal{S} & \text{if } Z_{min} = A \\ \gamma \dot{\gamma}_R \mathcal{S} & \text{if } Z_{min} = R \end{cases} \tag{6.34}$$

And for the relay zone as:

$$\mathbf{Z}_{\mathrm{R}} = \begin{cases} \dot{\gamma}_A / \gamma \, \mathcal{S} & \text{if } Z_{min} = A \\ \dot{\gamma}_R \mathcal{S} & \text{if } Z_{min} = R \end{cases} \tag{6.35}$$

6.5.7 Flow scheduling

Flows will be scheduled at BS based on an updated list mentioned earlier. The first entry in that list will be served before the others. The number of packets that will be scheduled from each queue will be dependent on the capacity of both access and relay slots allocated in the previous steps. The minimum number of scheduled bits for access traffic is equal to $\min(\alpha_a Z_A, W_a)$, where W_a is the backlogged access traffic for node a. To improve the efficiency, we maximize the utilization of resources. Therefore, we define a *utilization variable* for node a, $\mu_a(v)$ which can be calculated as

$$\mu_a(v) = \min_{\forall j \in \mathfrak{F}(a)} \left(\mathbf{Z}_R \left(\tau(h_j) \right) \left(\beta_{g_R} \left(\tau(h_j) \right) - \beta_j \left(\tau(h_j) \right) \right) q_j - \tilde{Q}_j \right) \tag{6.36}$$

where $\tau(h_j)$ is the application time which is equal to $v - (h_{max} - h_j) + 1$, and $\tilde{Q}_j$ is the most recent value of waiting traffic for node j at its father RS; we will use the latest reported queue size.

Note that each RS is not required to report the average relay link quality q_j on every scheduling cycle, as these links are fixed. For proper operation, the BS performs

route maintenance to check if there is a broken relay link in the cell. During this route maintenance operation, the RS's are required to report their link qualities. The idea is that we estimate relay link quality to enhance allocation. More frequent updates of $q_i$can increase the utilization at the cost of increased overhead. In fact, the algorithm will work even without this information. We can set the quality of each relay link to be at maximum. The updated queue size will then regulate the amount of traffic that can pass through that link. One of the reasons for having the queue size information included is to enhance the stability of relay queues.

The data size of scheduled packets of node a denoted by $D_a(v)$ is given by

$$D_a(v) = \min\big(\mu_a(v) + \alpha_a(\tau(h_a))Z_A(\tau(h_a)) - Q_{A_a}, \alpha_{g_A}(\tau(h_a))Z_A(\tau(h_a)) - Q_{A_a}, W_a\big) \tag{6.37}$$

Again, we have included the queue size for access traffic to assure relay queue stability. The number of available resources may increase or decrease at relay links due to the adaptation of $D_a(v)$. Thus, the amount of scheduled traffic can also be changed. Note that radio resource allocation does not change. What changes is the amount of traffic that can utilize the allocated resources. The reduction or the increase in available resources will be denoted by $\tilde{u}_a(v)$ and is expressed as:

$$\tilde{u}_a(v) = D_a(v) - \alpha_a(\tau(h_a))Z_A(\tau(h_a)) \tag{6.38}$$

This change in utilization may affect the utilization of other AN traffic. This effect translates in changing the values β_j of every father and child node of AE a. We

denote the set of nodes that is a father or child of AE a by $\mathcal{N}(a) = \mathfrak{F}(a) + \mathfrak{C}(a)$, where $\mathfrak{C}(a)$ is the set of children nodes of AE a. The modified β_j will be

$$\beta_j\left(\tau(h_j)\right) = \beta_j\left(\tau(h_j)\right) + \frac{\tilde{u}_a(v)}{\mathbf{Z}_A(\tau(h_a))q_j} \quad , j \in \mathcal{N}(a) \tag{6.39}$$

At the beginning of each scheduling cycle, a new β_j is calculated using Eq. 6.21. This will be increased or decreased within a scheduling cycle, using Eq. 6.41.

At each intermediate node j, the relay traffic is received along with its relay link allocation using Eq. 6.32. The relay will forward the traffic in the designated resources, i.e., the relay traffic is centrally scheduled and allocates the resources. However, AMC mode selection and queue reporting are done at RS's. This operation is done with minimum feedback requirements.

For access traffic Equation 6.33 gives the maximum allocated slots for access traffic of an access node. The access node will perform a PF index calculation after each allocation. The highest PF index flow is reported to the BS, which will update the BS flow list. Upon arrival of new traffic, the RS will use the previously calculated PF index as a ranking criterion by which the flows are scheduled.

We should distinguish between *resource allocation* and *flow scheduling*. In this algorithm, resource allocation for relay links as well as access links assigned to a relay is not required to be updated every scheduling instance. We may define an allocation interval as a constitution of multiple frames. On the other hand, scheduling is the

process by which a flow is selected for scheduling and packets are allocated at every frame. This implies that the scheduling interval equals the duration of one frame.

We present below the pseudo-code for the procedures of our A-DSGFS algorithm.

Algorithm 6.2 DSGFS Algorithm

(a) AE and Zone Resource Allocation at time v

% The time index,v, of the following variables is removed

Calculate Q_y; for all AN.

Calculate $Q_a = \sum_{y \in AE_a} Q_y$; for all AE.

For all group g_A^*DO

$m_{g_A^*} = \max_{a \in g_A^*}(Q_a)$;

$\alpha_{g_A^*} = \frac{m_{g_A^*}}{\sum_{\forall g_A} m_{g_A}}$;

END DO

$\alpha_y = \frac{Q_y}{Q_a}\alpha_{g_A(y)}$; for all AN.

$\beta_y = \frac{\alpha_y q_{A_y}}{q_y} + \sum_{\forall h \in \mathcal{R}(y)} \frac{\alpha_h q_{A_h}}{q_y}$; for all AN.

$\beta_a = \sum_{y \in AE_a} \beta_y$; for all AE.

$\beta_{g_R} = \max_{\forall a \in g_R}(\beta_a)$;for all group g_R.

$\gamma = \sum_{\forall g_R} \beta_{g_R}$

$Z_A = \frac{1}{\gamma + 1} S$

$Z_R = \frac{\gamma}{\gamma + 1} S$

$s_a = s_{g_R(a)} = \beta_{g_R(a)} Z_R / \gamma$; Allocation for Relay Zone

$\mathfrak{S}_a = \mathfrak{S}_{g_A(a)} = \alpha_{g_A(a)} Z_A$; Allocation for Access Zone

(b) Flow Scheduling for AE a at the BS at time v

For all AE a DO

$\mu_a(v) = \min_{\forall j \in \mathcal{F}(a)} \left(Z_R\left(\tau(h_j)\right)\left(\beta_{g_R(a)}\left(\tau(h_j)\right) - \beta_j\left(\tau(h_j)\right)\right) q_j - \tilde{Q}_j\right)$;

where $\tau(h_j) = v - (h_{max} - h_j) + 1$.

$D_a(v) = \min(\mu_a(v) + \alpha_a(\tau(h_a)) Z_A(\tau(h_a)) - Q_{A_a}, \alpha_{g_A}(\tau(h_a)) Z_A(\tau(h_a)) - Q_{A_a}, W_a)$;

$\tilde{u}_a(v) = D_a(v) - \alpha_a(\tau(h_a)) Z_A(\tau(h_a))$;

$\beta_j = \beta_j + \frac{\tilde{u}_a(v)}{Z_A(\tau(h_a)) q_j}$; $\forall j \in \mathcal{R}(a)$.

END DO

For all AN y DO

$s_y = \frac{\beta_y}{\beta_a} s_a$; Allocation for Relay Zone

$\mathfrak{S}_y = \frac{Q_y}{Q_a} \mathfrak{S}_a$; Allocation for Access Zone

$D_y(v) = \frac{Q_y}{Q_a} D_a(v)$; Traffic scheduled for AN y

END DO

(c) Resource allocation at AN y at time v

f_ranked_{v-1}; Assume order list of flows, if it is not available the order will be arbitrary
$Count = 0$; Initialize the counter
While $\mathfrak{G}_y > 0$ & $Count < \hat{u}(y)$; Note that $\mathfrak{G}_y = \mathfrak{G}_y(v - h_{max} + 1)$, where h_{max} is the maximum hop count
$Count = Count + 1$; $i^* = f\_rank_{v-1}(Count)$; The flow is selected based on the flow list
$D_{i^*}(v) = \min(W_{i^*}(v), q_{i^*}\mathfrak{G}_y)$; $W_i^*(v)$ is the traffic queued at AN y, q_i^* is the link quality of flow i^*.
$\hat{x}_{i^*}(v) = \frac{D_{i^*}(v)}{q_{i^*}}$; The number of resources allocated for scheduled flow is calculated.
$W_{i^*}(v) = W_{i^*}(v-1) - D_{i^*}(v) + B_{i^*}(v)$; $B_{i^*}(v)$ is the new traffic for flow i^* arrived at AN y
END While
$f\_rank_v = SORT(\{i\}, \{U_i(v)\})$; $SORT$ is a sort procedure to rank flows according to $U_i(v)$
$f\_rank_y(1)$ is reported and stored at the BS flow list.

(d) Flow scheduling at the BS

For all AN DO
Initialize $\mathbf{D}_i(v) = 0$; %the scheduled traffic of each flow is initialized by zero value
$Count = 0$;
While $\sum_{i \in u(y)} \mathbf{D}_{i^*}(v) < D_y(v)$ & $Count < \hat{u}(y)$
$Count = Count + 1$;
$i^* = f\_list_y(Count)$; %access the topmost entry in the flow list
$\mathbf{D}_{i^*}(v) = \min(Q_{i^*}(v-1), D_y(v))$; The scheduled traffic is limited by queue size and allocated traffic of AN a.
$Q_{i^*}(k) = Q_{i^*}(k-1) - \mathbf{D}_{i^*}(k) + \mathbf{B}_{i^*}(k)$; The Queue of the flow is updated. $\mathbf{B}_{i^*}(k)$ is the new arrival traffic.
If $Q_{i^*}(k) = 0$
RH_Rotate(f_list_y); to move the topmost entry of the list to the bottom when the queue is empty
End
END While
RANK_Update(f_list_y, i^*); Update the ranking of the flows with the new top ranked flow
END DO

6.6 Performance Simulation

6.6.1 Simulation Models and Parameters

As per our knowledge, there are no scalable resource allocation algorithms devised for multi-hop relay OFDMA cellular networks that are able to support IRRR. For this reason, we compare our algorithms with the distributed algorithm proposed in [60].

The same system parameters and assumptions used in the previous chapter are repeated here for convenience. MATLAB is used as our simulation environment to test throughput and fairness performance. Table 3-1 presents the parameters. We performed at least 20 experiments with 50 drops per experiment and 200 frames per drop. The large scale channel impairments (shadowing and path loss) were fixed during each experiment. A network containing 19 cells was assumed. We studied the performance in the central cell. The nodes inside the center cell experienced a continuous interference from the surrounding cells caused by "full queue" transmission of the interfering nodes.

A frequency reuse of one was assumed, which meant that each cell used all radio resources. The cells were partitioned into three sectors and the available resources and the number of users was distributed equally on each sector. We considered two examples for relay deployments: two-hop deployment that contains six one-hop RS's (Figure 6.2-b) and a three-hop deployment containing a total of 18 RS's, 6 one-hop RS's and 12 two-hop RS's (Figure 6.2-a). Tables 5.1 and 5.2 show

our groupings as discussed in section 5.7 for DSGFS algorithm, whereas Table 6.1 shows the grouping of the DSQS system.

A NLOS link was assumed between an access node and its subordinate UT's, while a LOS link was assumed between a node and another node if they had a direct father-child relationship; otherwise, NLOS link was assumed. The LOS modified IEEE 802.16 type-D path loss model was used for the BS/RS to RS links with 3.4dB shadowing if they had a fatherhood relationship. The remaining links were assumed NLOS, which used the modified IEEE 802.16 path loss model type B with 9.6 dB shadowing standard deviation. [14]

SUI 1 model was used for the multipath fading channel connecting the RS's with other RS's and with the BS. For the links to UT's, we considered the ITU Ped-A channel model. Refer to Table 3.1 for the other model assumptions.

6.6.2 Results and Discussions

Throughput and fairness performance comparison between our algorithms and the modified version of the distributed algorithm are presented in reference system [60]. As we mentioned earlier, the algorithm suffers from low utilization if traffic load is low since the algorithm relies on the RS's in scheduling the flow that can be admitted, assuming a full queue traffic model. The RS executes PF utility maximization to select the best candidate flow. This scheduling decision is sent back to the father node, which may admit or reject that flow schedule. The scheduling goes on until it reaches the BS, which in turn forwards the traffic of the scheduled flow. No

solution is provided if there is no traffic for the selected flow. In this case, the scheduled radio resources will be wasted.

We adopted a simple modification to this algorithm by letting the BS replace the scheduled flow by another flow that belongs to the same access node. We refer to the modified version of the algorithm as the Distributed Proportional Fair (DPF) algorithm.

Figure 6.4 and Figure 6.5 display the throughput performance of our algorithms compared to the DPF algorithm for the cell with 72 UT's and 18 UT's respectively. The figures show the two relaying cases mentioned earlier: two-hop and three-hop relaying. In each case, we show the average throughput vs. different users' demand including full-queue cases.

While reuse capability improved the DSQS compared to the DPF, the absence of selective relayed flow scheduling improved the DSQS over the DPF for the two-hop cases by 20%. While this may be considered as remarkable, it is *not* as significant as the A-DSGFS which registered an improvement of over 50% (Figure 6.4-a and Figure 6.5-a). This shows the significance of selective flow scheduling based on utility maximization in enhancing system throughput. This gain is usually referred to as *multi-user diversity* gain. User diversity gain is absent in DSQS, which results in reduced improvement compared to A-DSFGS. However, due to greater enhanced utilization, DSQS was able to perform much better than DPF. In cases of three-hops, the DSQS improvement over DPF increased by a further 44% because the DPF suffered from increased contention in system resources as hop count increased. This

was not the case in the DSQS and the A-DSFGS which possessed enhanced resource availability due to their IRRR capability.

A-DSGFS can select the best candidate flow and enjoys enhanced throughput performance due to better user diversity. Compared to DSQS, an improvement of approximately 20% is considered significant, especially in two-hop cases. The A-DSGFS registers a larger improvement compared to DPF due to the enhanced radio reuse capability coupled with multi-user diversity, especially in three-hop cases. More than 77% throughput improvement was recorded for the three-hop case over DPF.

If we do not allow for adaptive zone allocation, we should expect a reduction in throughput performance. We showed the impact of static zone allocation by providing an example of Static-DSGFS (S-DSGFS) scenario when we assigned the resources statically, assuming a uniform user distribution and same link quality at all hops. With this assumption, one third of the resources would need to be assigned for access traffic and two thirds assigned for relay in three-hop relaying cases. While for two-hop cases, the resources would be split in two halves; one half for the access traffic, the other half for relay traffic. This is not the best allocation strategy for S-DSGFS. In fact, we could optimize the allocation to match the user distribution and their link quality, which in turn would perform close to the adaptive system, especially under a full-queue traffic assumption. The purpose of displaying such an example for S-DSGFS is to emphasize the importance of tuning resource allocation to better match traffic and channel conditions. Compared to the adaptive case, static allocation caused an average of 20% throughput reduction.

The number of users has its own impact on performance, as it affects the demand on cell resources. In addition, link grouping showed some sensitivity to the number of users. Increasing the number of users caused a small improvement in throughput. Note that we do not consider the MAC overhead in our calculations. If it is considered, increasing the number of users will have a negative impact on throughput. Our algorithm improves with increasing the number of users, which may reduce throughput reduction caused by increased MAC overhead.

Fairness has been depicted in Figure 6.6 and Figure 6.7, for the cell with 72 and 18 hops respectively. Fairness is affected mainly by outage performance and resource availability. Hence, three-hop cases are usually better than two-hop cases due to outage performance. In low demand, better fairness for DPF was recorded due to its better outage performance. As demand increased, increased resources availability due to IRRR capability made DSQS and A-DSGFS perform better than DPF. The A-DSGFS also showed better fairness compared to DSQS at moderate load conditions due to enhanced resources utilization from improving multi-user diversity. In such cases, users were satisfied in the DSGFS. Increasing the demand further caused the A-DSGFS to utilize more resources. Thus, we noticed the DSQS exhibiting better fairness performance. This greedy behavior caused degraded fairness performance in the A-DSGFS compared to DPF at higher load conditions. The enhanced utilization procedure implemented in the A-DSGFS tried to allocate resources as much as possible to UT's, even if it would violate the fairness condition, whereas the multi-user diversity resulting in some UT's tended to have higher throughput.

The CDF of average user throughput is presented in Figure 6.8 and Figure 6.9 for the aforementioned two relaying scenarios. Considering the two-hop cases , the throughput improvement of our algorithms was apparent in all cases except in Figure 6.8-a, which was due to very low load experienced by the cell; the advantage is for DPF better SINR outage performance. At a higher load, Figure 6.8-b, c and d showed the advantage of IRRR. In Figure 6.8-b, we notice that the A-DSGFS was able to deliver the required demand for 70% of the users while the DPF guaranteed only 0.75Mbps. On the other hand, the DSQS was able to guarantee 1.45Mbps for the same number of users. In Figure 6.8-c, we present the case when the number of users is 72 with 0.5Mbps average demand. More than 75% percent of users had their demand fulfilled in the A-DSGFS case, while the DPF could not guarantee more than 0.18Mbps for the same number of users. We also observed that more than 0.3Mbps was guaranteed by DSQS for approximately the same number of users. Figure 6.8-d shows the greedy behavior of A-DSFGS due to its method for improving utilization, unlike the DSQS, which exhibits a fair behavior at high load conditions.

Figure 6.9 depicts the performance of three-hop cases. A-DSGFS delivers better throughput in all scenarios yet again. Figure 6.9-a shows the same performance in all cases. Figure 6.9-b, however, shows that 70% of UT's had their demand served by the A-DSGFS. The DSQS also performed well compared to the DPF algorithm with more than 1.4Mbps delivered to the same number of users, while only 1.1Mbps could be delivered in the DPF case.

The same observation can be seen in Figure 6.9-c, as it is effectively the same as Figure 6.9-b, since they have the same network load. Figure 6.9-d shows the performance at high demand. The average users' throughput of all algorithms is always less than the demand. We had seen that the DSQS maintained fairness by providing almost the same throughput for 90% of the users whereas A-DSFGS could not achieve such fairness due to its greedy behavior, where a relatively large number of users exceeded the average user throughput.

Average user throughput vs. distance has been shown in Figure 6.10 and Figure 6.11 for three-hop and two-hop cases, respectively. Our algorithms constantly show an improved user throughput in all cases except in Figure 6.10-a. As the load becomes higher, the superiority of our system becomes more apparent, especially the A-DSFGS system. This high throughput is a result of improved resource utilization. Our algorithms maintained a flat curve in all cases because of adopting a fair strategy in allocating the resources, which was beneficial in having ubiquitous service coverage throughout the cell. On the other hand, whenever the load exceeded cell capacity, the A-DSGFS algorithm performed as a greedy scheduler because of its method of increasing utilization, which may possibly lead to a decrease in algorithm fairness. However, at normal traffic conditions, A-DSFGS was capable of delivering better coverage compared to other algorithms.

6.7 Conclusion

In this chapter, we have proposed two decentralized scheduling algorithms devised for OFDMA relay enhanced cellular networks: DSQS and DSFGS. The DSQS was proposed to provide efficient flow scheduling and resource allocation with minimum interaction from the BS. The DSQS showed good throughput and fairness performance compared to the reference system.

The DSFGS, on the other hand, had further improved the throughput performance due to its better user diversity. We implemented UT listing at the BS to store and update the best candidate nominated by the RS to which the nominated UT belonged. This enhanced multi-user diversity, which improved resource utilization and throughput performance. Implementing a generalized fairness principle, the operator may select an appropriate merit by which the resources are distributed over the AN's. In our performance evaluation, we have used PF merit for allocating RS resources and PF flow index to select the appropriate flow.

IRRR was supported using link grouping. Enhanced throughput has been achieved by both the DSQS and the DSGFS. The improvement was higher in the cases when the DSGFS was deployed due to its better resources utilization and multi-user diversity. Both algorithms are of linear complexity and require insignificant overhead. Thus, they are scalable and can be used for larger networks and larger hop counts.

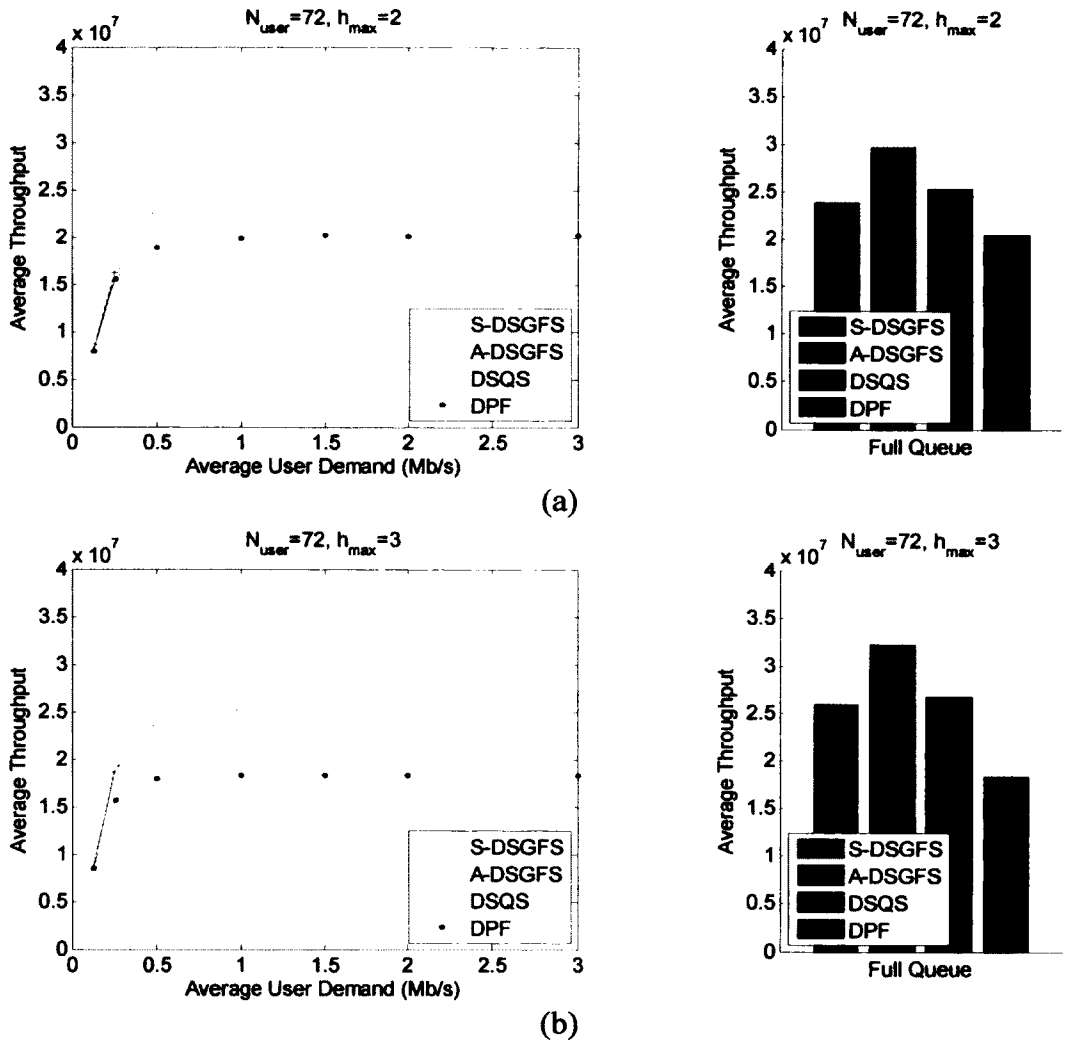


Figure 6.4 Average Cell Throughput vs. users demand for the cell with 72 users, and with different number of hops: (a) two-hop relaying, (b) three-hop relaying.

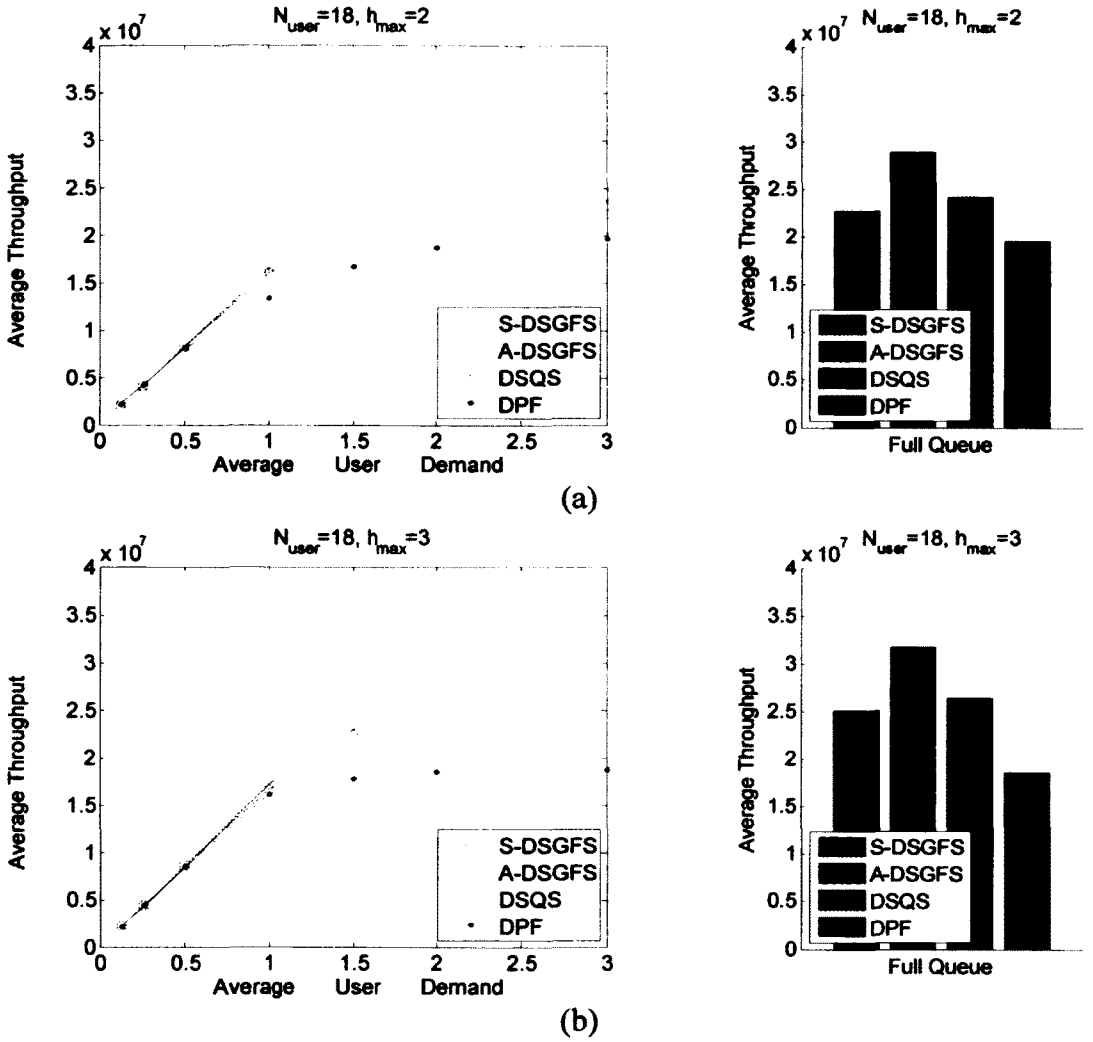
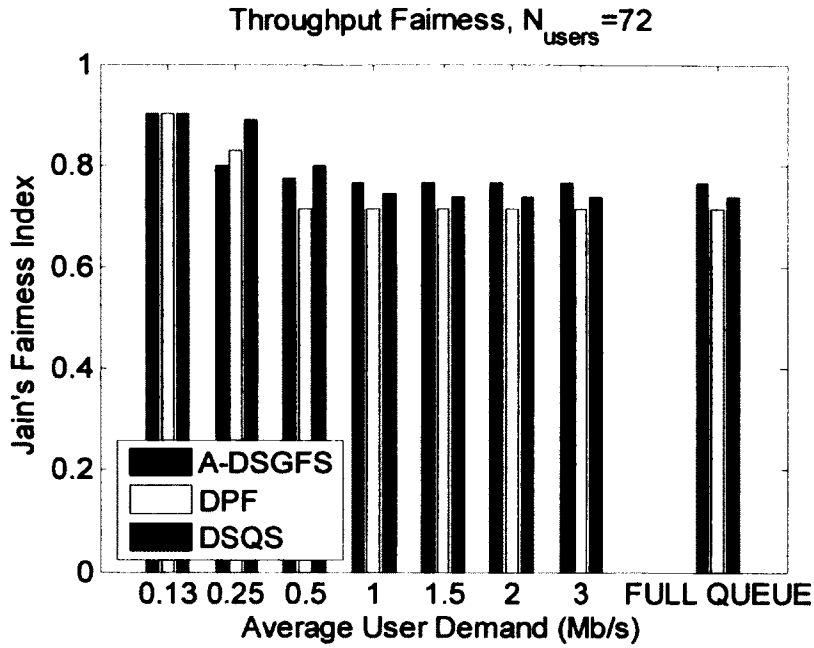
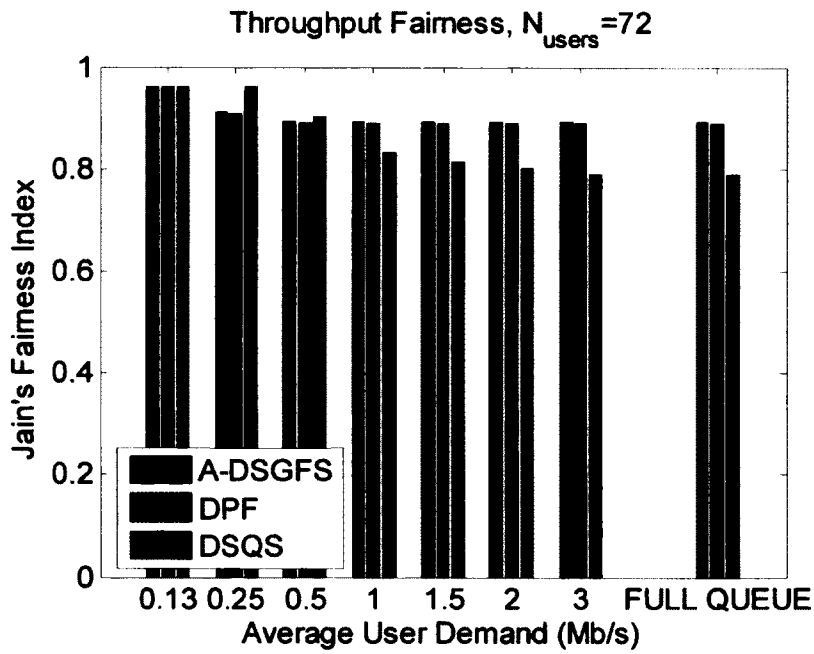


Figure 6.5 Average Cell Throughput vs. users demand for the cell with 18 users, and with different number of hops: (a) two-hop relaying, (b) three-hop relaying.

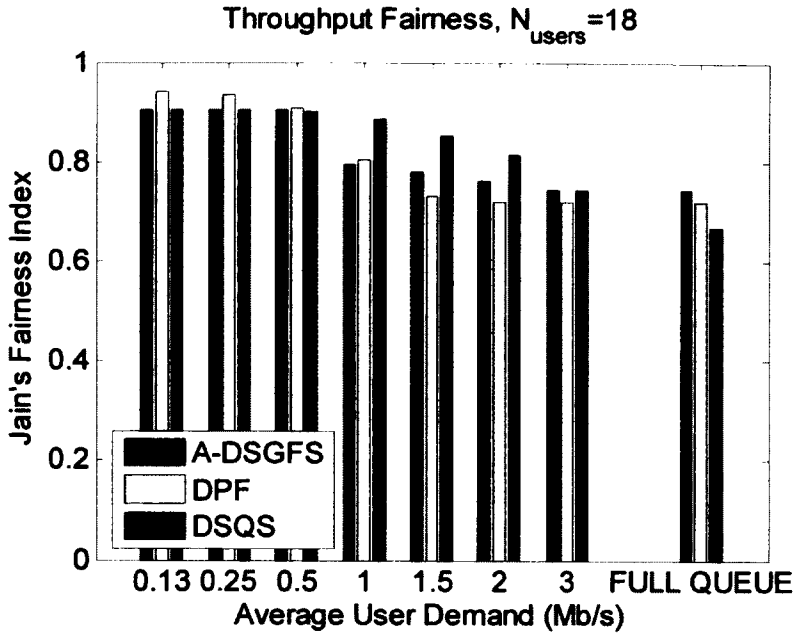


(a)

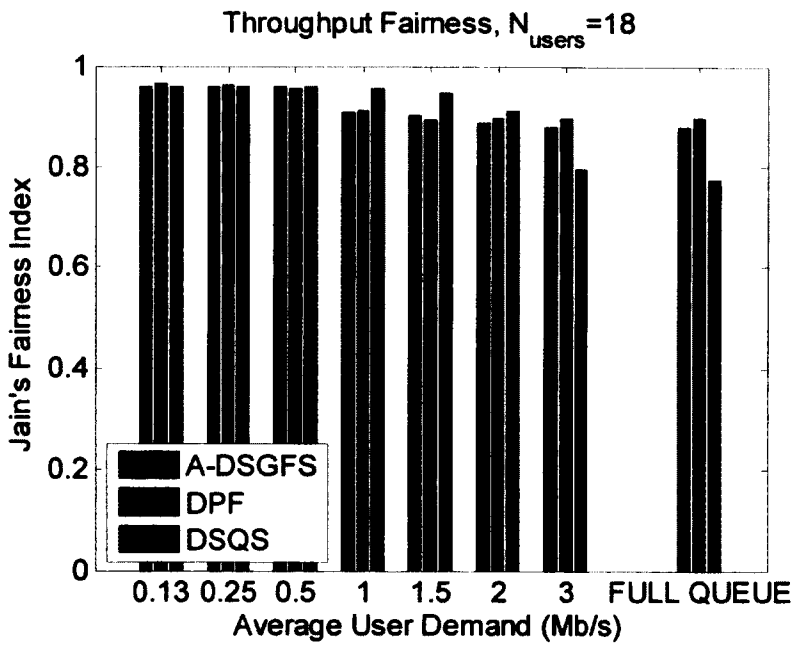


(b)

Figure 6.6 Average Fairness vs. users demand for the cell with 72 users, and with different number of hops: (a) two-hop relaying, (b) three-hop relaying.

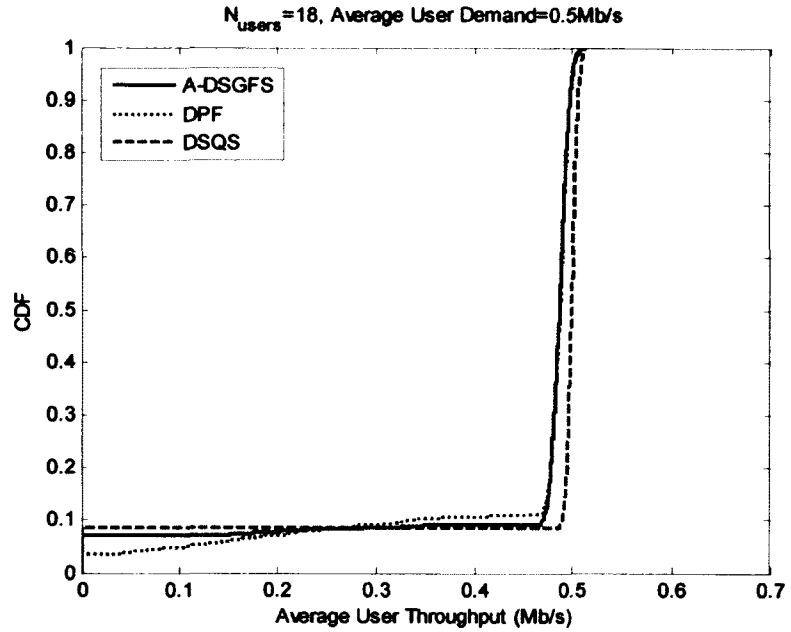


(a)

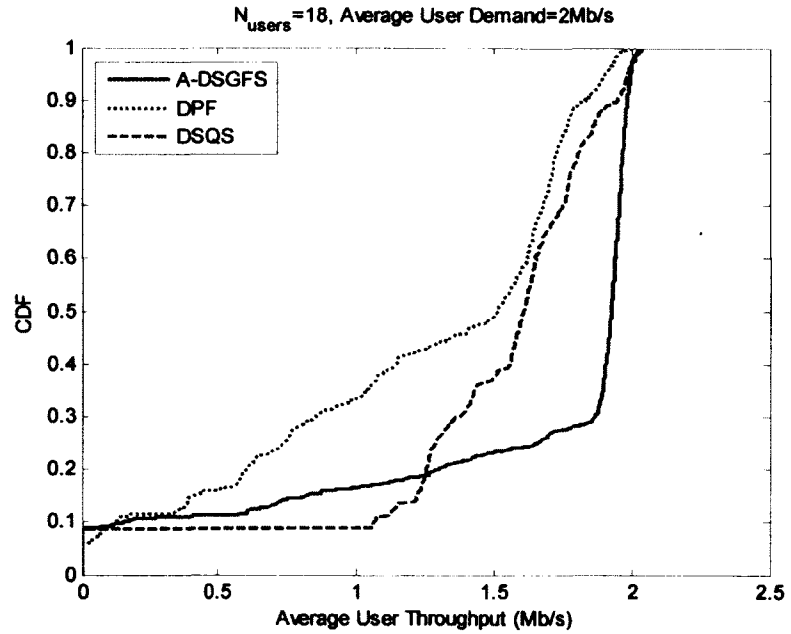


(b)

Figure 6.7 Average Fairness vs. users demand for the cell with 18 users, and with different number of hops: (a) two-hop relaying, (b) three-hop relaying.

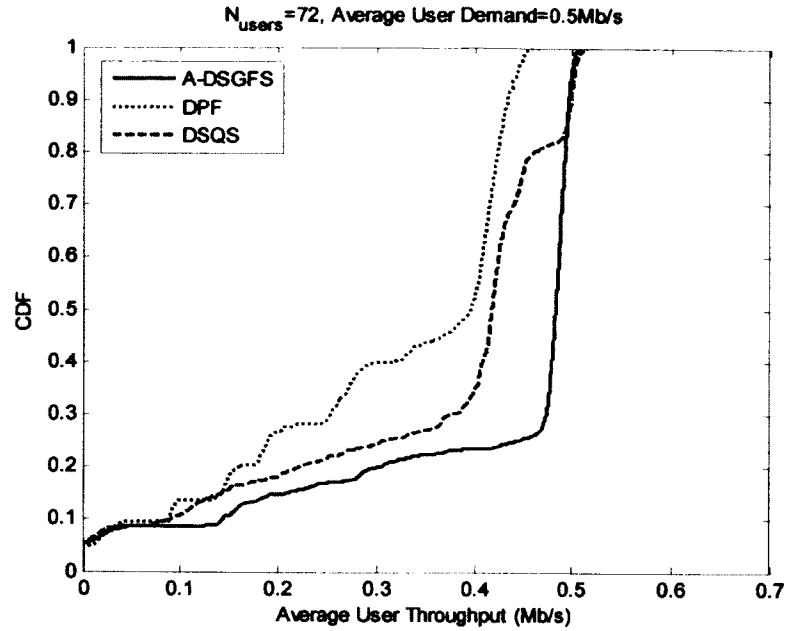


(a)

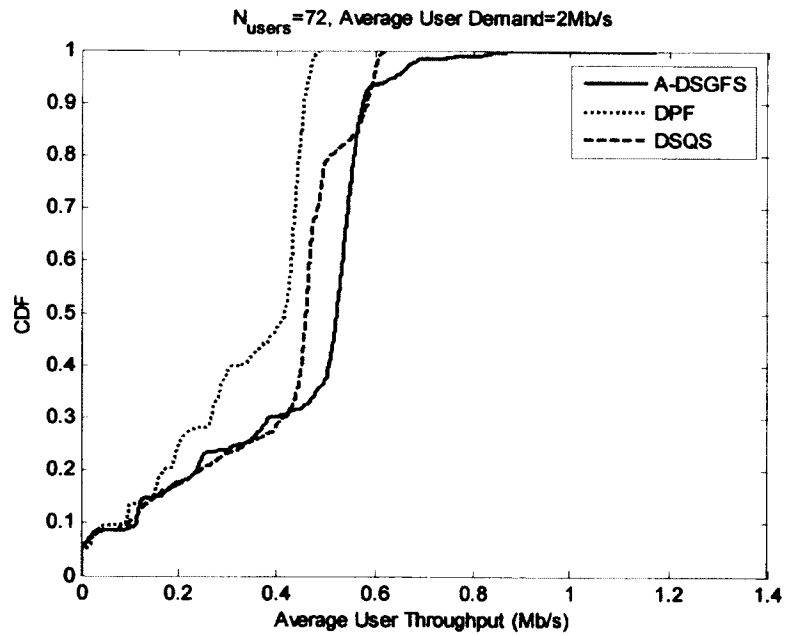


(b)

Figure 6.8 CDF of Average Users throughput of our algorithms considering two-hop cell layout: (a) N_{users}= 18, average demand equals 0.5Mb/s (b) N_{users}= 18, average demand equals 2Mb/s, , (c) N_{users}= 72, average demand equals 0.5Mb/s (d) N_{users}= 72, average demand equals 2Mb/s.

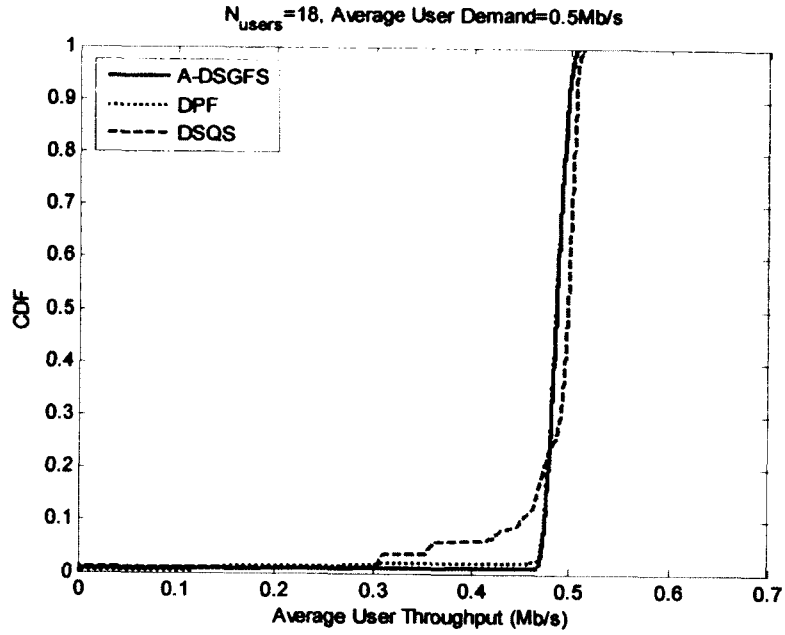


(c)

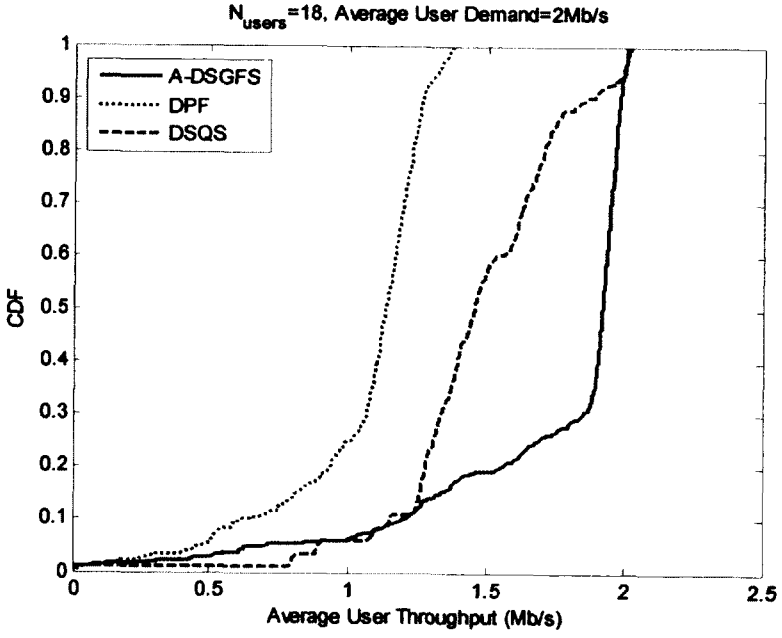


(d)

Figure 6.8 (Cont.) CDF of Average Users throughput of our algorithms considering two-hop cell layout: (a) N_{users}= 18, average demand equals 0.5Mb/s (b) N_{users}= 18, average demand equals 2Mb/s, , (c) N_{users}= 72, average demand equals 0.5Mb/s (d) N_{users}= 72, average demand equals 2Mb/s.

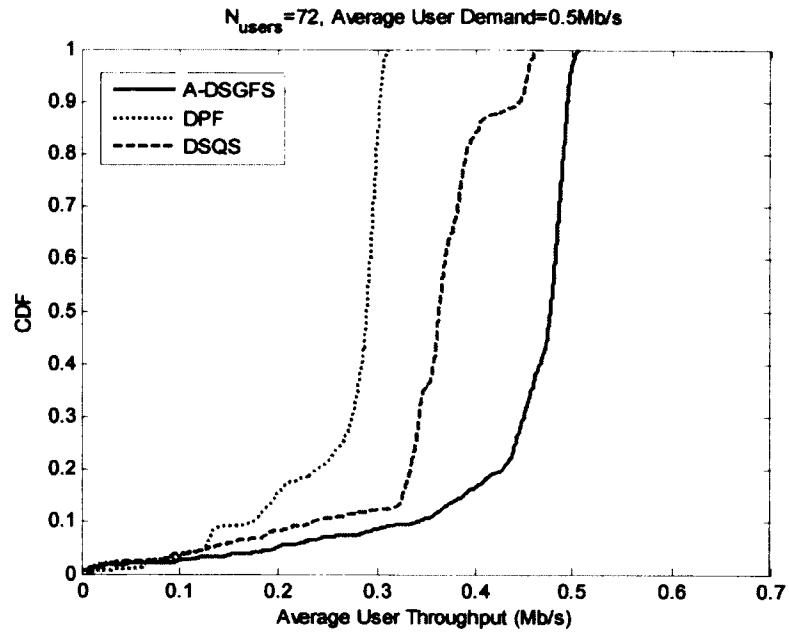


(a)

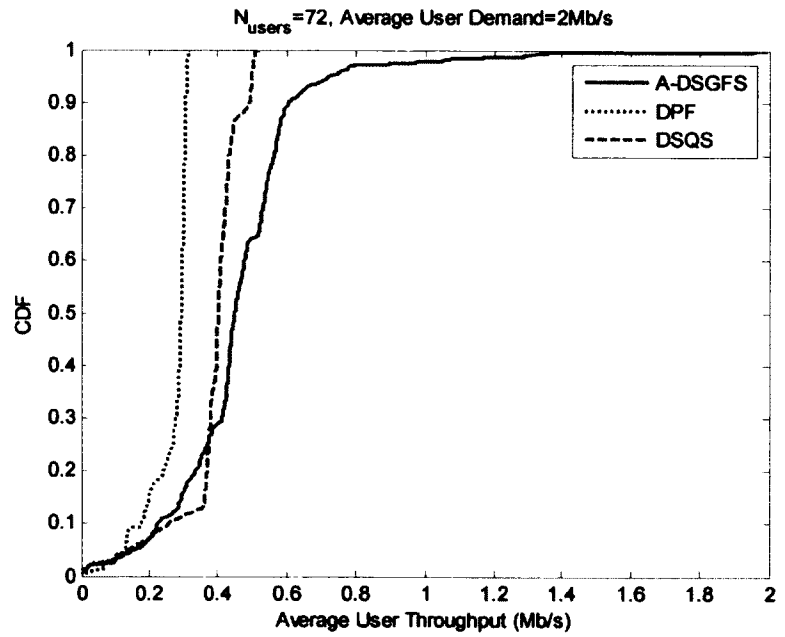


(b)

Figure 6.9 CDF of Average Users throughput of our algorithms considering three-hop cell layout: (a) N_{users}= 18, average demand equals 0.5Mb/s (b) N_{users}= 18, average demand equals 2Mb/s, , (c) N_{users}= 72, average demand equals 0.5Mb/s (d) N_{users}= 72, average demand equals 2Mb/s.

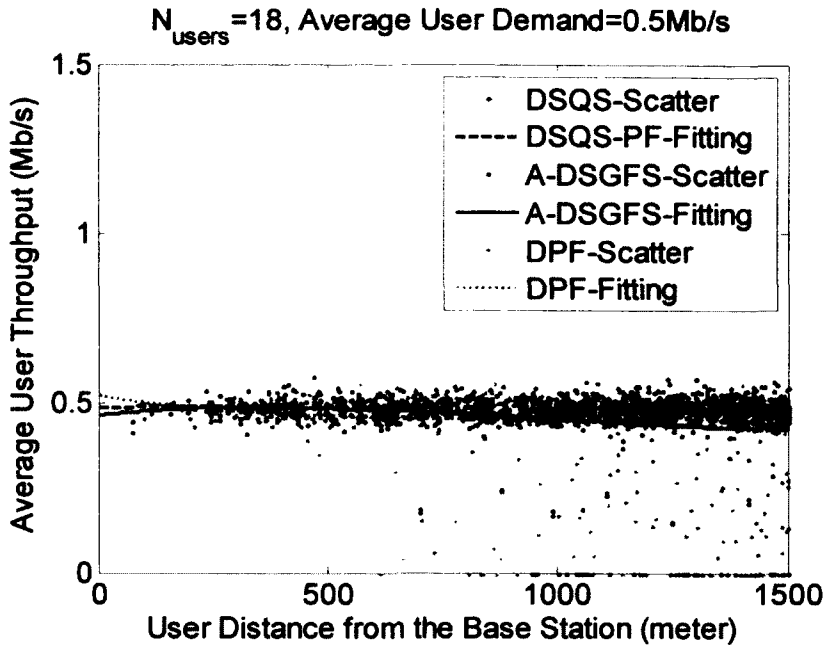


(c)

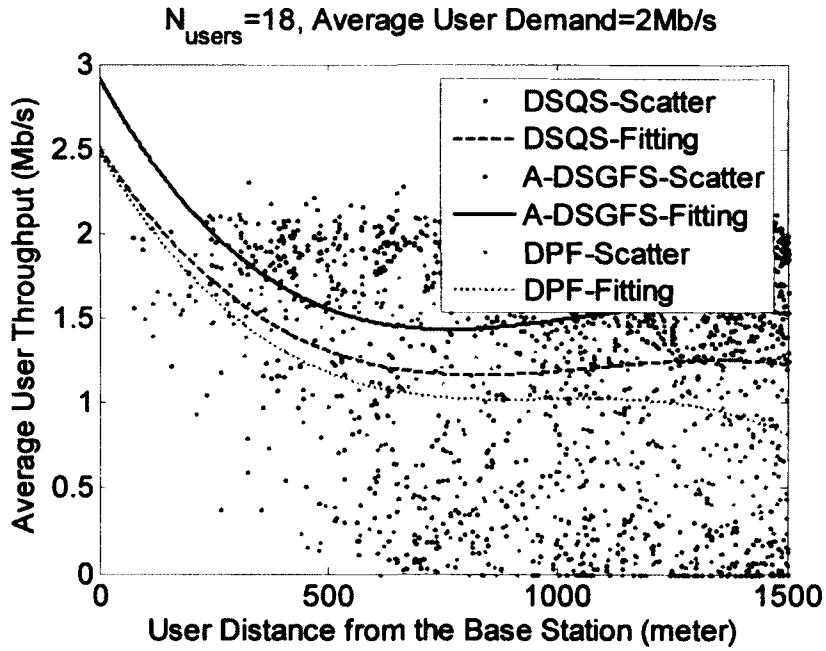


(d)

Figure 6.9 (Cont.) CDF of Average Users throughput of our algorithms considering three-hop cell layout: (a) N_{users}= 18, average demand equals 0.5Mb/s (b) N_{users}= 18, average demand equals 2Mb/s, , (c) N_{users}= 72, average demand equals 0.5Mb/s (d) N_{users}= 72, average demand equals 2Mb/s.

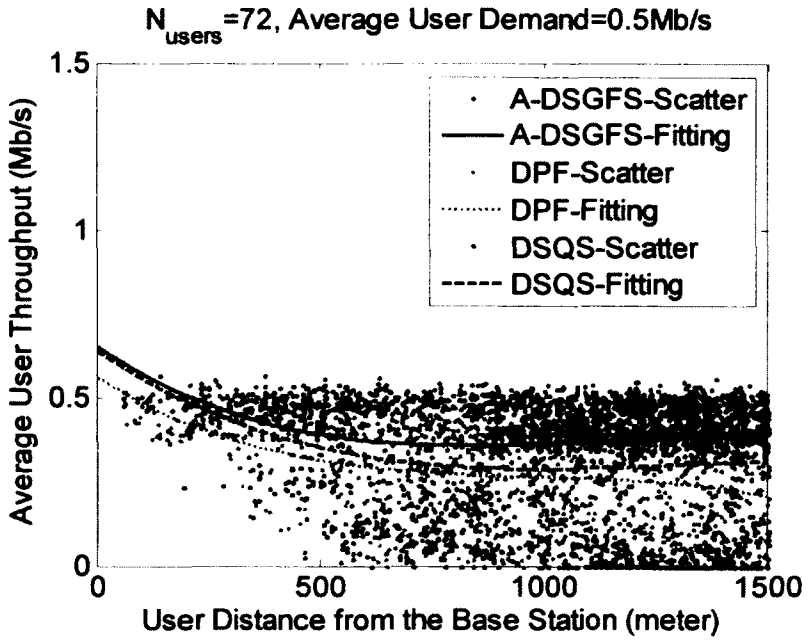


(a)

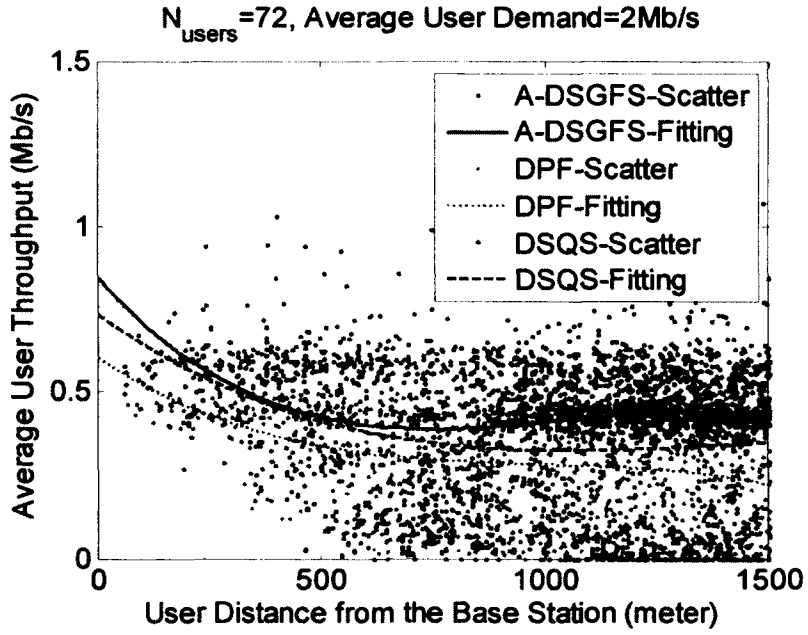


(b)

Figure 6.10 Scatter plot and fitting curves for Average Users throughput vs. distance for cells considering two-hop cell layout: (a) N_{users}= 18, average demand equals 0.5Mb/s (b) N_{users}= 18, average demand equals 2Mb/s, , (c) N_{users}= 72, average demand equals 0.5Mb/s (d) N_{users}= 72, average demand equals 2Mb/s.

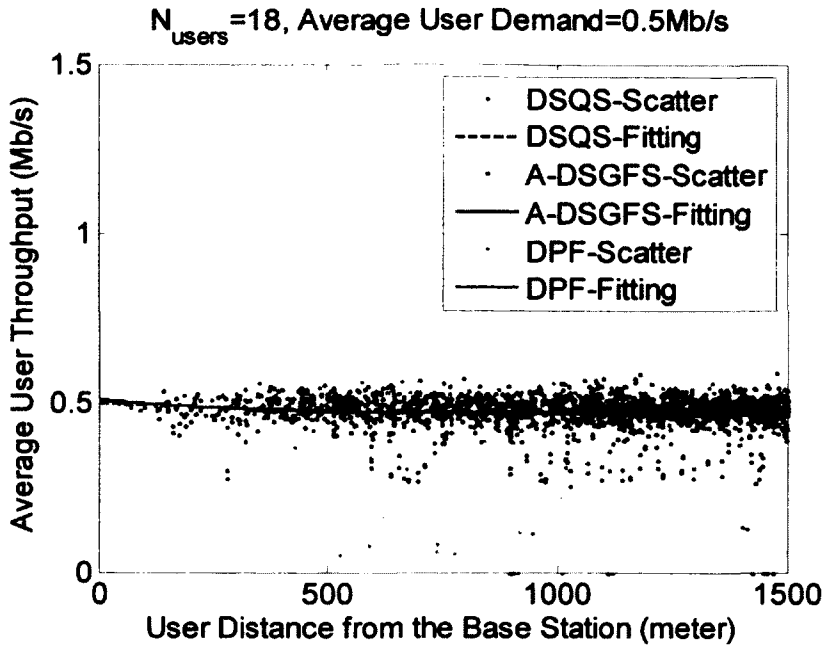


(c)

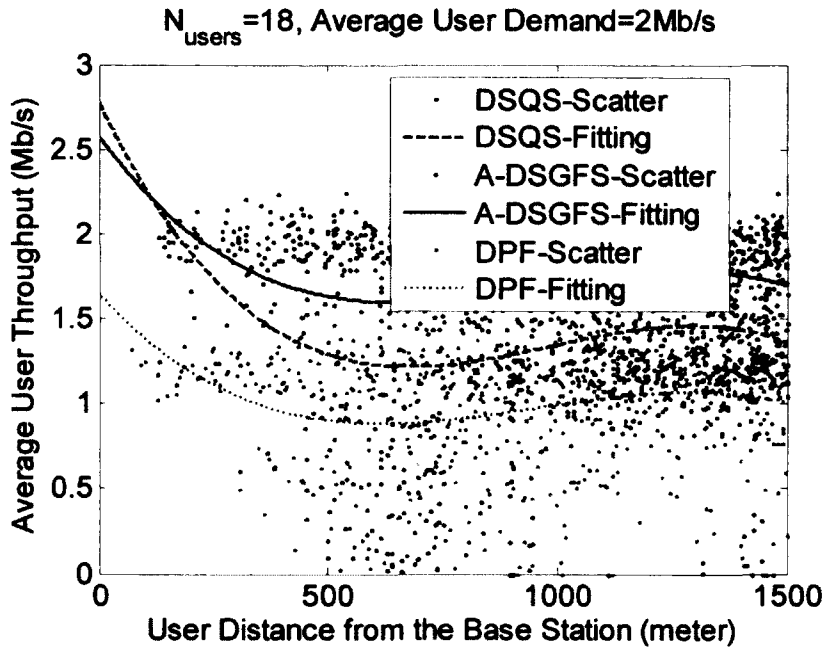


(d)

Figure 6.10 (Cont.) Scatter plot and fitting curves for Average Users throughput vs. distance for cells considering two-hop cell layout: (a) N_{users}= 18, average demand equals 0.5Mb/s (b) N_{users}= 18, average demand equals 2Mb/s, , (c) N_{users}= 72, average demand equals 0.5Mb/s (d) N_{users}= 72, average demand equals 2Mb/s.

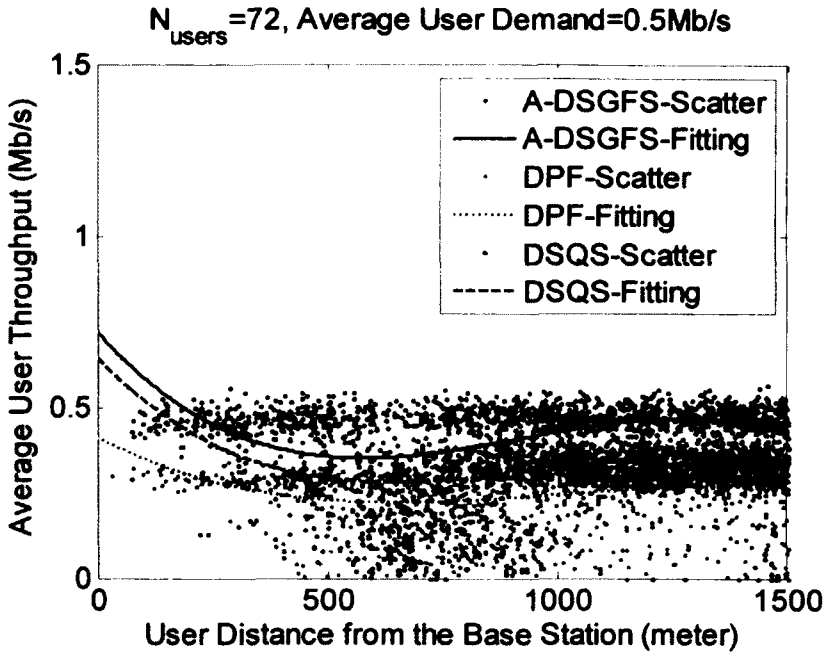


(a)

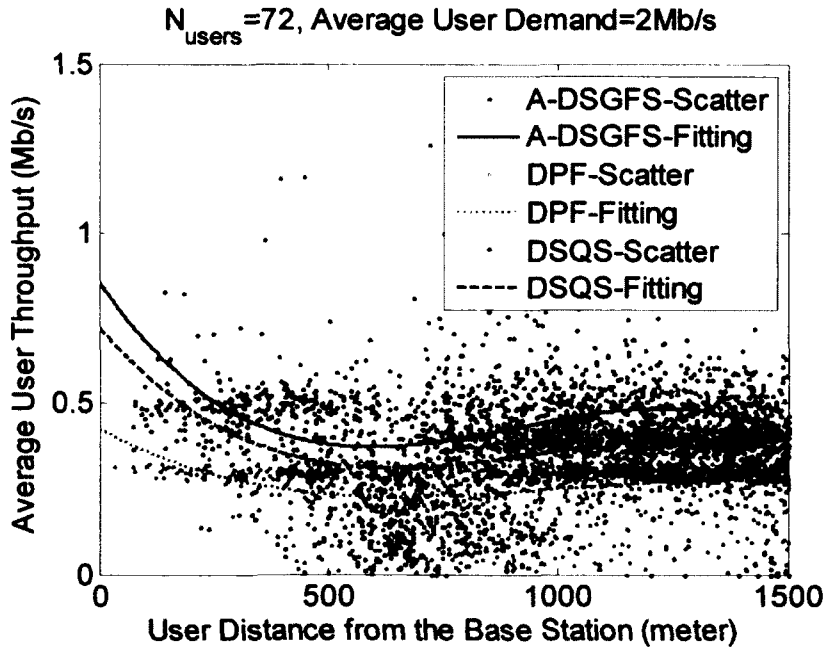


(b)

Figure 6.11 Scatter plot and fitting curves for Average Users throughput vs. distance for cells considering three-hop cell layout: (a) N_{users}= 18, average demand equals 0.5Mb/s (b) N_{users}= 18, average demand equals 2Mb/s, , (c) N_{users}= 72, average demand equals 0.5Mb/s (d) N_{users}= 72, average demand equals 2Mb/s.



(c)



(d)

Figure 6.11 (Cont.) Scatter plot and fitting curves for Average Users throughput vs. distance for cells considering three-hop cell layout: (a) N_{users}= 18, average demand equals 0.5Mb/s (b) N_{users}= 18, average demand equals 2Mb/s, , (c) N_{users}= 72, average demand equals 0.5Mb/s (d) N_{users}= 72, average demand equals 2Mb/s.

Chapter 7

Conclusions and Future Work

7.1 Conclusions

In Chapter 3, we examined the spectral efficiency performance of different radio resources partitioning and reuse patterns for two and three-hop cells. We have also studied the impact of IRRR, number of hops, power control, and the selection of optimal AMC mode on spectral efficiency. We have found that IRRR is essential to enhance the capacity of cellular relay networks. In this case, we select the best AMC for each link to achieve the desired performance enhancement. If we sacrifice the IRRR capability in order to improve SIR coverage and to simplify the forwarding algorithm, our results show that by choosing the worst link AMC mode for all links, there is no significant degradation on throughput performance. We expect the reduced overhead, time and processing requirements to eventually result in better performance.

We have formulated the scheduling problem for OFDMA relay enhanced networks. Due to its complexity and NP-hardness, we proposed our approaches to simplify the problem with or without IRRR capability. The first approach was to adopt the multi-frame concept with no. The second is the partial IRRR which is an enhancement of the pervious approach by restricting the number of frames to two and

allow for a partial reuse of resources. This scheme improved scalability and throughput performance with a modest increase of complexity due to power control requirement. The third approach was the Full IRRR which was based on link grouping according to interference conditions. These three approaches are the basis of the algorithms proposed in chapters 4, 5 and 6.

In chapter 4, we proposed the WLMF algorithm following the first approach, aimed at simple algorithm with reduced overhead and relaying delay. It is based on the multi-frame concept where we restrict the activation of the relay links according to their hop count. Fixing the AMC mode and slot allocation of a flow over all hop links transfers the scheduling problem to a legacy (no-relay) scheduling problem. This algorithm did not result in a significant performance loss compared to the case where we chose the best AMC mode over all links. In fact, due to its low overhead requirement and lower processing delay, we expect to have improved performance in terms of throughput and delay.

WLDF is a modified version of WLMF proposed to improve the throughput and enhance the scalability of the algorithm. WLDF follows the second approach which restricts the multi-frame to two frames only regardless of the number of hops. It forces the relay links to reuse the resources on every second hop. With a simple control mechanism, the WLDF provided better throughput and enhanced utilization which led to a better scalable algorithm.

At high load, the WLMF suffered from reduced throughput due to the increased contention on radio resources. This is, in fact, not related to the scheduling algorithm

itself. Rather, it is related to relay deployment, routing, and selection algorithm. With no IRRR capability, the WLMF falls short quickly of radio resources to cope with the demand caused by the contention on system radio resources between relay links and access links. On the other hand, with lower network demand, we noticed improved performance because of the improved SINR coverage. WLDF improved the throughput performance by partial reuse of radio resources.

This suggests that WLMF can be a viable option for coverage enhancement of the cell if there are few users and/or users with low demand. This is because of its improved SINR coverage. On the contrary, for higher throughput, we improve the utilization of resources by enabling partial IRRR as in WLDF. The improvement can be increased further by enabling full IRRR potential, which is what we proposed in the next two chapters.

Algorithms in chapters 5 and 6 follow the third approach that relies on link grouping to enable full IRRR capability. In chapter 5, we proposed the TBSA, a centralized algorithm capable of supporting full IRRR. It adopted the concept of end-to-end tunneling to reduce the complexity of allocation process while being able to enhance the utilization by using tunnel metrics. The flow based allocation process was carried out with ease while preserving IRRR capabilities. IRRR is supported by link grouping; links are grouped such that the members of a group can be activated simultaneously without overwhelming each other.

TBSA in its static version (S-TBSA) was able to enhance the throughput compared to the no-reuse cases. The improvement became significant when we

adapted the resource allocation of link groups. We proposed the novel pre-allocation algorithm to adaptively tune group allocations. It provided the original algorithm (S-TBSA) with adapted zone and group allocations. This adaptive algorithm showed a remarkable enhancement of system throughput with a moderate increase in algorithm complexity.

In chapter 6, we proposed two de-centralized schemes that required reduced overhead and feedback: the DSQS and the DSFGS algorithms. The DSQS relied on queue monitoring and reporting of RS queue while the BS responded by allocating the requested resources. The relay grouping scheme was proposed to increase resource availability and hence reduce delays in relay queuing. It also enhances the utilization and, hence, improves the throughput. Fairness performance was also improved as a result of fair allocation procedures.

Multi-user diversity by means of a flow listing procedure has helped our DSGFS algorithm to improve the throughput. Adaptation of resource allocation for access nodes has shown a great impact on throughput performance. The improvement in throughput has been gained with insignificant overhead cost. Moreover, the complexity of those algorithms is linear, making them suitable for larger networks with a larger number of relay hops.

7.2 Future Work

7.2.1 Delay Performance

Multi-hopping increases delay as the nodes require time to receive the relayed data, process, and retransmit. The nodes may need additional time waiting for a scheduled opportunity to transmit. Clearly, time delay in the multi-hop relay network is very critical. Many aspects affect delay, such as traffic type and demand, number of users, scheduling algorithm, and objective function. In our work, we focused on throughput, and fairness performance of the scheduling algorithms we proposed. The extension to our work that analyzes system delay is of much value. System delay is affected not only by the algorithm we propose, but also by the type of carried traffic and delay limit. The burstiness of traffic plays a significant role in delay performance. The focus in this case should be on the design of the objective function, by which flows are selected such that a good tradeoff between delay reduction and throughput maximization can be attained. Our centralized algorithms WLMF, WLDF, and TBSA are built on flow selection mechanism, based on utility maximization. This utility function can be designed to enhance delay performance.

The delay in distributed cases is more problematic, since no flow based mechanisms exists to satisfy the per flow delay requirements. In such cases, the traffic has to be classified according to delay requirement, where more urgent flows are prioritized over the less urgent ones. Delay performance is critical, especially for real

time traffic. Relaying complicates the delay problem because of its variability, uncertainty, and relay queuing delays. Physical layer models may need further simplification in order to be able to analyze queuing performance at both the BS and the RS levels, especially if we want to develop solutions for delay sensitive applications.

7.2.2 HARQ Support

HARQ support in these systems is of great importance to provide a reliable service. Unfortunately, HARQ with relaying can be challenging, especially when delay requirements are stringent. We propose a simple solution by assigning a special channel that is common to all RS's for quick error recovery. This channel can be used only once per error event. This non-persistent use of the common channel is proposed to prevent bad links from continuously reusing the common channel, which may render it unusable. Our preliminary results show a promising improvement in relaying delay at a minimal resource cost.

7.2.3 Different Objectives

We have considered simple throughput based objective algorithms, which may be appropriate for best effort traffic. More fine-tuned objective functions may be implemented to support a certain class, or different classes of QoS. While our algorithm is traffic aware, some QoS requirement request specific delay limits and minimum throughput. Therefore, a flow utility function can be designed to support a class or multiple classes of QoS. The literature is rich in the area of QoS based allocation proposals for legacy systems, where different utility functions have been

proposed. The solutions proposed in the literature for legacy systems can be developed further under our scheduling frameworks.

Furthermore, scheduling problems can be formulated in a manner where the utility function is optimized under the constraint of average throughput being satisfied. This kind of formulation transfers the problem to a stochastic optimization problem. It is a worthwhile extension to develop a stochastic based solution under our proposed scheduling frameworks.

7.2.4 Mobility Support

Mobility requires frequent handover from one RS to another. Such handovers can cause some reduction in throughput performance, not only because of the overhead associated with such handovers, but also because of the waste resulting from dropping packets that are transmitted for relayed traffic as a result of changing the path. A remedy of such problems is to allow the UT to keep the connection of the old RS, as well as the new RS, and keep receiving from both during a specified transmission period.

The performance with mobility assumption needs to be assessed. A question in the case of mobility may be asked: is it worth all the complexity of relaying in the case of mobility, can the scheduler at the BS wait for some time hoping that the UT arrives at the favored spot, and then forward the traffic without going to the complexity of relaying? The answer of this question depends on many actors, such as the speed of

the UT, type of traffic, delay limit, location of the UT and RS's, which can be addressed.

This also suggests that a speed and location based relay selection algorithm may achieve an enhanced throughput and delay performance by letting the UT's associated with an RS not only based on throughput or SINR performance, but based on their speed and direction of movement. In our framework, the relay selection procedure is split from the allocation procedure and, as such, our solutions can be implemented with the aforementioned relay selection procedure.

In de-centrally controlled relays, there may be some packets that are queued in an AN or at one of the intermediate nodes. The intermediate nodes will forward all the packets to the AN according to the tunneled operation we have adopted. The AN will in turn have accumulated packets for the flow that has been handed over. The solution we mentioned earlier can help, which allows a transition period where the UT can receive from two access nodes. If the communication to the handed-over UT is lost, the AN either will drop the waiting packets, or an alternative remedy may be implemented. As an added solution, we suggest to allow the BS assign a temporary tunnel between the two access nodes that are associated with the handover, over which the old packets transmit the queued packets. The question of how to allocate the resources of this tunnel should be addressed, whether we allocate dedicated resources, or we get the resources from the link group to which this relay belongs. This procedure is not required in centralized solutions, since there are no relay queues.

7.2.5 MIMO and advance antenna configurations

MIMO and advanced antenna configurations are becoming the standard for future communication systems. They can provide enhanced throughput performance by increasing spectral efficiency caused by spatial multiplexing and/or transmit diversity gains. The AN decides first how the antenna configuration is used. It may be used for transmit diversity and hence increase the link quality. This option does not affect the scheduling mechanism, except for increasing the slot instantaneous throughput. The advanced antenna configuration may also be used for spatial multiplexing to enhance spectral efficiency by accepting more than one flow for the same radio resources. In our link grouping procedure, we can support the spatial multiplexing capability by adding spatially multiplexed links in the same link group. Thus, the scheduler considers the multiplexed links as different relays and the algorithms can be implemented without any extra complexity. In WLMF and WLDF, all links must have the same antenna configurations for spatial multiplexing to be considered.

7.2.6 Cooperative Relaying

Cooperative relaying is a mechanism that is used to allow the RS's and BS to collaborate in transmitting the data in order to achieve better spectral efficiency. Basically, the cooperated nodes will act as a distributed MIMO (or MISO) system. Relaying in its SISO configuration has a simple form of cooperation which involves selecting the best relay for transmitting the data. A more involved cooperation configuration can be adopted. WLMF and WLDF can be easily extended to support

cooperative operation by letting two or more relays form a relay group, which are allowed to transmit cooperatively to the same UT's. Implementing flow based allocation, although not optimal, allows a simple integration of cooperative relay operation.

Cooperation can be implemented in the other algorithms, too. It is better implemented in the handover procedure. A threshold on the received SINR can trigger the handover procedure, which results in assigning a new access node. The BS can transmit the same data to the old and the new access nodes. Both will use the same access resources allocation if the two access nodes belong to the same link group or use different resources if they belong to a different link group. In either case, transmit diversity can be achieved. It is likely to find the new AN sharing one or more relay links with the old access node. Therefore, it is appropriate to share the same links during the transition period. We propose to implement what we referred to as pivot tunneling, where the tunnel ends at an intermediate relay. Later, new tunnels between the intermediate RS and access nodes are established. At this point, a scheduling algorithm should be modified to support the new tunneling structure. We propose to use the same algorithms in TBSA and DSGFS, but the utilization metric calculation needs to be modified to account for the new tunnel structure. In DSGS, it will rely in routing tables to forward the traffic and a duplicate entry has to be allowed for the pivot RS to transmit to the two cooperative access nodes.

References

[1] Jeffrey G. Andrews, Arunabha Ghosh, and Rias Muhamed, *Fundamentals of WiMAX: Understanding Broadband Wireless Networking*. Prentice Hall, 2007.

[2] S.-J. Lin, W.-H. Sheen, I.-K. Fu, and C.-C. Huang, "Resource Scheduling with Directional Antennas for Multi-Hop Relay Networks in a Manhattan-Like Environment," in *Mobile WiMAX*, K.-C. Chen and J. R. B. de Marca, Eds. Chichester, UK: John Wiley & Sons, Ltd, 2008.

[3] F. I-Kang, S. Wern-Ho, and R. Fang-Ching, "Deployment and radio resource reuse in IEEE 802.16j multi-hop relay network in Manhattan-like environment," *in Proceedings, International Conference on Information, Communications & Signal Processing*, pp. 1–5, December 2007.

[4] H. Yanikomeroglu, D. D. Falconer, and V. Sreng, "Coverage enhancement through two-hop peer-to-peer relaying in cellular radio networks," World Wireless Research Forum meeting no. 7, Eindhoven, the Netherlands, December 2002.

[5] I. K. Chan and W. Liao, "Adaptive bandwidth allocation for TCP traffic in IEEE 802.16j wireless networks with transparent relay stations," *in Proceedings, IEEE 19th International Symposium on Personal, Indoor and Mobile Radio Communications*, pp. 1–5, September 2008.

[6] W. Nam, W. Chang, S.-Y. Chung, and Y. H. Lee, "Transmit Optimization for Relay-Based Cellular OFDMA Systems," *in Proceedings, IEEE International Conference on Communications*, pp. 5714–5719, June 2007.

[7] J. Han, J. Lv, T. Liang, and J. Zhang, "Joint routing and subcarrier allocation for OFDMA-based relay systems," *in Proceedings, Third International Conference on Communications and Networking in China*, pp. 589–594, 25 August 2008.

[8] B. Fan, W. Wang, Y. Lin, L. Huang, and K. Zheng, "Subcarrier Allocation for OFDMA Relay Networks with Proportional Fair Constraint," *in Proceedings, IEEE International Conference on Communications*, pp. 1–5, June 2009.

[9] S. Deb, V. Mhatre, and V. Ramaiyan, "WiMAX relay networks: opportunistic scheduling to exploit multiuser diversity and frequency selectivity," *in Proceedings, Proceedings of the 14th ACM international conference on Mobile computing and networking*, pp. 163-174, 2008.

[10] O. Oyman, "OFDM2A: A Centralized Resource Allocation Policy for Cellular Multi-hop Networks," *in Proceedings, Fortieth Asilomar Conference on Signals, Systems and Computers*, pp. 656–660, October 2006.

[11] M. Salem, A. Adinoyi, M. Rahman, H. Yanikomeroglu, D. Falconer, and Young-Doo Kim, "Fairness-aware radio resource management in downlink OFDMA cellular relay networks," *IEEE Transactions on Wireless Communications*, vol. 9, no. 5, pp. 1628-1639, 2010.

[12] Liping Wang, Yusheng Ji, Fuqiang Liu, and Jie Li, "Performance Improvement through Relay-Channel Partitioning and Reuse in OFDMA Multihop Cellular Networks," *International Wireless Communications and Mobile Computing Conference (IWCMC '08)*, pp. 177-182, August 2008.

[13] Matthew Baker, "LTE-Advanced Physical Layer," Beijing, 17 December 2009. [Online]. Available: http://www.3gpp.org/ftp/workshop/2009-12-17_ITU-R_IMT-Adv_eval/docs/pdf/REV-090003-r1.pdf

[14] M. Hart and J. J. Son, "Multi-hop Relay System Evaluation Methodology (Channel Model and Performance Metric)," 2007. [Online]. Available: www.ieee802.org/16/relay/docs/80216j-06_013r3.pdf.

[15] "IEEE Standard for Local and Metropolitan Area Networks Part 16: Air Interface for Fixed and Mobile Broadband Wireless Access Systems Amendment 2: Physical and Medium Access Control Layers for Combined Fixed and Mobile Operation in Licensed Bands and Corri," 2006. [Online]. Available: http://ieeexplore.ieee.org/xpl/freeabs_all.jsp?arnumber=1603394.

[16] J. N. Laneman, D. N. . Tse, and G. W. Wornell, "Cooperative Diversity in Wireless Networks: Efficient Protocols and Outage Behavior," *IEEE Transactions on Information Theory*, vol. 50, no. 12, pp. 3062–3080, December 2004.

[17] M. Yuksel and E. Erkip, "Diversity in relaying protocols with amplify and forward," *in Proceedings, IEEE Global Telecommunications Conference (GLOBECOM '03)*, pp. 2025–2029, December 2003.

[18] Su Chang Chae, "Direct Relaying Zone in Frame Structure for Transparent Mode." September 2007.

[19] Hasna, M.O. and Alouini, M.-S., "A performance study of dual-hop transmissions with fixed gain relays," *IEEE Trans. Wireless Commun.*, vol. 3, no. 6, November 2004.

[20] A. Ribeiro and G. B. Giannakis, "Symbol error probabilities for general Cooperative links," *IEEE Transactions on Wireless Communications*, vol. 4, no. 3, pp. 1264–1273, May 2005.

[21] F. H. P. Fitzek and M. D. Katz, Eds., *Cooperation in Wireless Networks: Principles and Applications*, 2nd ed. Dordrecht, Netherlands: Springer, 2006.

[22] R. S. Cheng, K. B. Lataief, and R. D. Murch, "Multiuser OFDM with adaptive subcarrier, bit, and power allocation," *IEEE Journal on Selected Areas in Communications*, vol. 17, no. 10, pp. 1747–1758, 1999.

[23] W. Rhee and J. M. Cioffi, "Increase in capacity of multiuser OFDM system using dynamic subchannel allocation," *in Proceedings, IEEE 51st Vehicular Technology Conference (VTC 2000-Spring)*, vol. 2, pp. 1085–1089, May 2000.

[24] K. B. Letaief, "Optimizing power and resource management for multiuser MIMO/OFDM systems," *in Proceedings, IEEE Global Telecommunications Conference, 2003 (GLOBECOM '03)*, pp. 179–183, 01 December 2003.

[25] M. Ergen, S. Coleri, and P. Varaiya, "Qos aware adaptive resource allocation techniques for fair scheduling in ofdma based broadband wireless access systems," *IEEE Transactions on Broadcasting*, vol. 49, no. 4, pp. 362–370, December 2003.

[26] P. Svedman, S. K. Wilson, L. J. Cimini, and B. Ottersten, "Opportunistic Beamforming and Scheduling for OFDMA Systems," *IEEE Transactions on Communications*, vol. 55, no. 5, pp. 941–952, May 2007.

[27] F. Kelly, "Charging and rate control for elastic traffic," *European transactions on Telecommunications*, vol. 8, no. 1, pp. 33–37, 1997.

[28] G. Barriac and J. Holtzman, "Introducing delay sensitivity into the proportional fair algorithm for CDMA downlink scheduling," *in Proceedings, IEEE Seventh International Symposium on Spread Spectrum Techniques and Applications,*, vol. 3, pp. 652–656, December 2002.

[29] K. Kuenyoung, K. Hoon, and H. Youngnam, "A proportionally fair scheduling algorithm with QoS and priority in 1xEV-DO," *in Proceedings, The 13th IEEE International Symposium on Personal, Indoor and Mobile Radio Communications*, vol. 5, pp. 2239–2243, September 2002.

[30] A. Wang, L. Xiao, S. Zhou, X. Xu, and Y. Yao, "Dynamic resource management in the fourth generation wireless systems," *in Proceedings, International Conference on Communication Technology Proceedings (ICCT 2003)*, pp. 1095–1098, April 2003.

[31] L. Badia, M. Lindstrom, J. Zander, and M. Zorzi, "Demand and pricing effects on the radio resource allocation of multimedia communication systems," *in Proceedings, IEEE Global Telecommunications Conference (GLOBECOM '03)*, pp. 4116–4121, 01 December 2003.

[32] H. J. Zhu, "Scheduling algorithms for ofdm broadband wireless systems," Ph.D. Thesis, Carleton University, Ottawa, ON, 2006.

[33] G. Song and Li, Y., "Adaptive subcarrier and power allocation in OFDM based on maximizing utility," *in Proceedings, The 57th IEEE Semiannual Vehicular Technology Conference (VTC 2003-Spring)*, vol. 2, pp. 905 - 909, April 2003.

[34] N. Enderle and X. Lagrange, "User satisfaction models and scheduling algorithms for packet-switched services in UMTS," *in Proceedings, The 57th IEEE Semiannual Vehicular Technology Conference (VTC 2003-Spring.)*, vol.3, pp. 1704–1709, July 2003.

[35] H. Mason, N. K. Shankaranarayanan, and P. Henry, "A subjective survey of user experience for data applications for future cellular wireless networks," *in Proceedings, Symposium on Applications and the Internet*, pp. 167–175, January 2001.

[36] I. C. Wong and B. L. Evans, "Optimal OFDMA Resource Allocation with Linear Complexity to Maximize Ergodic Weighted Sum Capacity," *in Proceedings, IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP '07)*, p. III–601–III–604, April 2007.

[37] V. Shankarkumar and N. H. Vaidya, "Medium access control protocols using directional antennas in ad hoc networks," *in Proceedings, Nineteenth Annual Joint Conference of the IEEE Computer and Communications Societies (INFOCOM 2000)*, pp. 13–21, March 2000.

[38] J. Li, C. Blake, D. S. J. De Couto, H. I. Lee, and R. Morris, "Capacity of Ad Hoc wireless networks," *in Proceedings, the 7th annual international conference on Mobile computing and networking (MobiCom '01)*, pp. 61–69, July 2001.

[39] B. S. Manoj and K. J. Kumar, "On the use of multiple hops in next generation wireless systems," *Wireless Networks*, vol. 12, no. 2, pp. 199 – 221, 2006.

[40] H. Wu, C. Qiao, De, S., and Tonguz, O., "Integrated cellular and ad hoc relaying systems: iCAR," *IEEE Journal on Selected Areas in Communications*, vol. 19, no. 10, pp. 2105–2115, October 2001.

[41] H.-Y. Hsieh and R. Sivakumar, "Performance comparison of cellular andmulti-hop wireless networks: a quantitative study.," *ACM Sigmetrics*, June 2001.

[42] R. Ananthapadmanabha, B. S. Manoj, and C. S. . Murthy, "Multi-hop cellular networks: the architecture and routing protocols," *In Proceedings, 12th IEEE International Symposium on Personal, Indoor and Mobile Radio Communications. (PIMRC 2001)*, p. G–78–G–82, September 2001.

[43] L. Ying-Dar and H. Yu-Ching, "Multihop cellular: a new architecture for wireless communications," *in Proceedings, IEEE Conference on Computer Communications. Nineteenth Annual Joint Conference of the IEEE Computer and Communications Societies (INFOCOM 2000)*, pp. 1273–1282, March 2000.

[44] Mukherjee, S. and Viswanathan, H., "Resource allocation strategies for linear symmetric wireless networks with relays," *in Proceedings, IEEE ICC2002*, vol. 1, pp. 366 - 370, August 2002.

[45] N. Challa and H. Cam, "Cost-aware downlink scheduling of shared channels for cellular networks with relays," *IEEE International Conference on Performance, Computing, and Communications*, pp. 793–798, 2004.

[46] H. Viswanathan and S. Mukherjee, "Performance of cellular networks with relays and centralized scheduling," *IEEE Transactions on Wireless Communications*, vol. 4, no. 5, pp. 2318–2328, September 2005.

[47] M. Charafeddine, O. Oyman, and S. Sandhu, "System-Level Performance of Cellular Multihop Relaying with Multiuser Scheduling," *in Proceedings, 41st Annual Conference on Information Sciences and Systems*, pp. 631–636, March 2007.

[48] I. Hammerstrom and A. Wittneben, "Temporal fairness enhanced scheduling for cooperative relaying networks in low mobility fading environments," *in Proceedings, IEEE 6th Workshop on Signal Processing Advances in Wireless Communications, 2005*, pp. 525–529, June 2005.

[49] I. Hammerstrom, M. Kuhn, and A. Wittneben, "Channel adaptive scheduling for cooperative relay networks," *in Proceedings, IEEE 60th Vehicular Technology Conference (VTC2004-Fall)*, pp. 2784–2788, 2004.

[50] D. H. Lee, S. C. Kim, D. C. Park, S. S. Hwang, and Y.-il Kim, "Performance evaluation of multi-hop relay system with deployment scenarios," *IEEE Military Communications Conference, 2008 (MILCOM 2008)*, pp. 1–7, 2008.

[51] S.-yeon Kim, S.-jin Kim, S.-wan Ryu, H.-woo Lee, and C.-ho Cho, "Performance analysis of single-frame mode and multi-frame mode in IEEE 802.16j MMR system," *in Proceedings, IEEE 19th International Symposium on Personal, Indoor and Mobile Radio Communications, 2008 (PIMRC 2008)*, pp. 1–5, September 2008.

[52] V. Genc, S. Murphy, Y. Yu, and J. Murphy, "IEEE 802.16J relay-based wireless access networks: an overview [recent advances and evolution of WLAN and WMAN standards]," *Wireless Communications, IEEE*, vol. 15, no. 5, pp. 56-63, 2008.

[53] L. Erwu, W. Dongyao, L. Jimin, Shen Gang, and Jin Shan, "Performance Evaluation of Bandwidth Allocation in 802.16j Mobile Multi-Hop Relay Networks," *in Proceedings, Vehicular Technology Conference (VTC2007-Spring). IEEE 65th*, pp. 939-943, 2007.

[54] L. Cuthbert, "User fairness analysis of a game theory based power allocation scheme in OFDMA relay systems," *European Wireless Conference*, pp. 173–177, May 2009.

[55] B. H. Walke, R. Pabst, and D. Schultz, "A mobile broadband system based on fixed wireless routers," *in Proceedings, International Conference on Communication Technology Proceedings 2003 (ICCT 2003)*, pp. 1310–1317, April 2003.

[56] W. Mohr, R. Luder, and K.-H. Mohrmann, "Data rate estimates, range calculations and spectrum demand for new elements of systems beyond IMT-2000," *in Proceedings, The 5th International Symposium on Wireless Personal Multimedia Communications*, vol. 1, pp. 37–46, 27 October 2002.

[57] Y.-N. Lee, J.-C. Chen, Y.-C. Wang, and J.-T. Chen, "A Novel Distributed Scheduling Algorithm for Downlink Relay Networks," *IEEE Transactions on Wireless Communications*, vol. 6, no. 6, pp. 1985–1991, Jun. 2007.

[58] M. Kaneko and P. Popovski, "Adaptive Resource Allocation in Cellular OFDMA System with Multiple Relay Stations," *in Proceedings, Vehicular Technology Conference, 2007. VTC2007-Spring. IEEE 65th*, pp. 3026-3030, 2007.

[59] L. Wang, Y. Ji, and F. Liu, "A Semi-Distributed Resource Allocation Scheme for OFDMA Relay-Enhanced Downlink Systems," *in Proceedings, IEEE Globecom Workshops*, pp. 1–6, November 2008.

[60] Ying Wang, Gen Li, Tong Wu, and Feng Gong, "Adaptive Proportional Fair Scheduling in Multihop OFDMA Systems," *Vehicular Technology Conference (VTC 2010-Spring), 2010 IEEE 71st*, pp. 1-5, 2010.

[61] L. Wang, Y. Ji, and F. Liu, "Joint Optimization for Proportional Fairness in OFDMA Relay-Enhanced Cellular Networks," *in Proceedings, IEEE Wireless Communication and Networking Conference (WCNC 2010)*, pp. 1–6, April 2010.

[62] Y. Wang, Z. Wu, H. Song, M. Peng, and W. Wang, "Centralized Spatial Reuse Designing for Fixed Multihop Relay Wireless Access Networks," *in Proceedings, International Conference on Communications, Circuits and Systems*, pp. 1230–1234, June 2006.

[63] C. Hoymann, K. Klagges, and M. Schinnenburg, "Multihop Communication in Relay Enhanced IEEE 802.16 Networks," *in Proceedings, IEEE 17th International Symposium on Personal, Indoor and Mobile Radio Communications, 2006*, pp. 1–4, September 2006.

[64] Z. Tao, A. Li, K. H. Teo, and J. Zhang, "Frame Structure Design for IEEE 802.16j Mobile Multihop Relay (MMR) Networks," *in Proceedings, IEEE Global Telecommunications Conference (GLOBECOM '07)*, November 2007.

[65] J. Z. Tao, K. H. Teo, J. Zhang, and T. Kuze, "An adaptive frame structure for OFDMA-based mobile multi-hop relay networks," 2007. [Online]. Available: www.ieee802.org/16/relay/contrib/C80216j-07_117.pdf.

[66] D. H. Ahn, J. Hui, K. H. Lee, and C.-wook Suh, "Multi-frame structure consistent to 802.16e for MR Networks," 2007. [Online]. Available: http://www.ieee802.org/16/relay/contrib/C80216j-07_162r5.doc.

[67] P. LI, M. RONG, Y. XUE, D. YU, L. WANG, and H. SHI, "Spectrum partitioning and relay positioning for cellular system enhanced with two-hop fixed relay nodes," *IEICE TRANSACTIONS on Communications*, vol. E90-B, no. 11, pp. 3181–3188, Mar. 2007.

[68] Jemin Lee, Sungsoo Park, Hano Wang, and Daesik Hong, "QoS-guaranteed Transmission Scheme Selection for OFDMA Multi-hop Cellular Networks," *IEEE International Conference on Communications*, pp. 4587-4591, June 2007.

[69] Won-Hyoung Park and Saewoong Bahk, "WLC25-3: Resource Management Policies for Fixed Relays in Cellular Networks," *in Proceedings, IEEE Global Telecommunications Conference (GLOBECOM '06)*, pp. 1-5, 27-November 2006.

[70] Ping Li, Mengtian Rong, Yisheng Xue, and Egon Schulz, "Reuse One Frequency Planning for Two-hop Cellular System with Fixed Relay Nodes," *in Proceedings, Wireless Communications and Networking Conference (WCNC 2007)*, pp. 2253-2258, March 2007.

[71] S. F. Meko and P. Chaporkar, "Channel partitioning and relay placement in multi-hop cellular networks," *in Proceedings, 6th International Symposium on Wireless Communication Systems (ISWCS 2009)*, pp. 66-70, September 2009.

[72] Shiang-Jiun Lin, Wern-Ho Sheen, and Chia-Chi Huang, "Downlink Performance and Optimization of Relay-Assisted Cellular Networks," *in Proceedings, IEEE Wireless Communications and Networking Conference (WCNC 2009)*, pp. 1-6, April 2009.

[73] Yue Zhao, Xuming Fang, Xiaopeng Hu, Zhengguang Zhao, and Yan Long, "Fractional frequency reuse schemes and performance evaluation for OFDMA multi-hop cellular networks," *in Proceedings, 5th International Conference on Testbeds and Research Infrastructures for the Development of Networks & Communities and Workshops (TridentCom 2009)*, pp. 1-5, April 2009.

[74] Zhangchao Ma, Kan Zheng, Wenbo Wang, and Yang Liu, "Route Selection Strategies in Cellular Networks with Two-Hop Relaying," *in Proceedings, the 5th International Conference on Wireless Communications, Networking and Mobile Computing (WiCom '09)*, pp. 1-4, September 2009.

[75] Z. Zhao, X. Fang, Y. Zhu, and Y. Long, "Two Frequency Reuse Schemes in OFDMA-TDD Based Two-Hop Relay Networks," *In Proceedings, IEEE Wireless Communications and Networking Conference,* pp. 1-6, April 2010.

[76] Wern-Ho Sheen, Shiang-Jiun Lin, and Chia-Chi Huang, "Downlink Optimization and Performance of Relay-Assisted Cellular Networks in Multicell Environments," *IEEE Transactions on Vehicular Technology*, vol. 59, no. 5, pp. 2529-2542, June 2010.

[77] M. Salem, A. Adinoyi, H. Yanikomeroglu, and Young-Doo Kim, "Nomadic Relay-Directed Joint Power and Subchannel Allocation in OFDMA-Based Cellular Fixed Relay Networks," *in Proceedings, IEEE 71st Vehicular Technology Conference (VTC 2010-Spring)*, pp. 1-5, May 2010.

[78] C. G. Kang, J. M. Ku, P. K. Kim, S. J. Lee, and S. Shin, "On the Performance of Broadband Mobile Internet Access System," *Wireless Personal Communications*, vol. 47, no. 2, pp. 265-279, 2008.

[79] J. Hui et al., "MR-BS and RS frame management for multi-frame structure," 2007. [Online]. Available: www.ieee802.org/16/relay/contrib/C80216j-07_450_2.doc.

[80] V. Sreng, H. Yanikomeroglu, and D. D. Falconer, "Relayer selection strategies in cellular networks with peer-to-peer relaying," *in Proceedings, IEEE 58th Vehicular Technology Conference*, vol. 3, pp. 1949-1953, August 2003.

[81] M. Salem et al., "An Overview of Radio Resource Management in Relay-Enhanced OFDMA-Based Networks," *Communications Surveys & Tutorials, IEEE*, vol. 12, no. 3, pp. 422-438, 2010.

[82] Junkai Zhang, Suili Feng, We Ye, and Hongcheng Zhuang, "MAC Performance Evaluation of IEEE 802.16j," *in Proceedings, International Symposium on Information Science and Engineering (ISISE '08)*, vol. 1, pp. 421-425, December 2008.

[83] M. Kaneko and P. Popovski, "Radio Resource Allocation Algorithm for Relay-Aided Cellular OFDMA System," *in Proceedings, IEEE International Conference on Communications, 2007 (ICC '07)*, pp. 4831–4836, June 2007.

[84] L. Wang, Y. Ji, and F. Liu, "A Semi-Distributed Resource Allocation Scheme for OFDMA Relay-Enhanced Downlink Systems," *IEEE Globecom Workshops*, pp. 1–6, November 2008.

[85] L. Wang, Y. Ji, and F. Liu, "Joint Optimization for Proportional Fairness in OFDMA Relay-Enhanced Cellular Networks," *IEEE Wireless Communication and Networking Conference (WCNC 2010)*, pp. 1–6, April 2010.

[86] Ingemar Kaj. 2002. *Stochastic Modeling in Broadband Communications Systems*. Soc. for Industrial and Applied Math., Philadelphia, PA, USA.

[87] D. Catterysse and L. V. Wassenhove, "A survey of algorithms for the generalized assignment problem," *European Journal of Operational Research*, vol. 60, no. 3, pp. 260–272, 1992.