Dynamic Resource Management for Next Generation Broadband Wireless Communication Systems

Thesis submitted in accordance with the requirements of the University of Liverpool for the degree of Doctor in Philosophy

by

Nan Zhou

September 2010

Declaration

The work in this thesis is based on research carried out at the University of Liverpool. No part of this thesis has been submitted elsewhere for any other degree or qualification and it is all my own work unless referenced to the contrary in the text.

Abstract

The next generation wireless communication systems are expected to provide high speed, high quality and multi-tasking services, which are supported by emerging technologies such as orthogonal frequency division multiplexing (OFDM) and co-operative relaying. Dynamic resource management, which is investigated in this thesis, plays an increasingly important role in system design of next generation broadband wireless systems, including the Long Term Evolution (LTE)/ World-wide Interoperability for Microwave Access (WiMAX) systems, and the relay based LTE-Advanced/WiMAX 2 systems.

In the first contribution, a cross-layer resource management scheme is proposed for the downlink LTE/WiMAX systems, to maximize the weighted sum capacity of multiple users. A packet dependent (PD) scheduling scheme is employed at the medium access control (MAC) layer, which determines the packet transmission order by assigning different weights to different packets, and is shown by simulations more efficient than previous methods where all packets in a queue have the same weight. The weight design in PD scheduling considers the delay, size and quality of service (QoS) priority level of packets. Each user weight employed in resource allocation at the physical (PHY) layer is obtained by summing up the weights of selected packets for the user. By properly choosing the number of packets used for weight calculation, the proposed cross-layer design requires a much lower overall complexity than the conventional queue based designs. It is also proven to achieve the maximum system stability region.

In the second contribution, an optimal resource allocation (RA) scheme is proposed for the downlink LTE-Advanced/WiMAX 2 systems, where a relay station is deployed to help the base station communicate with multiple out-of-range users (destinations), rather than a single destination as assumed in most previous work. The proposed RA scheme consists of subchannel allocation (SA) to multiple users, subcarrier pairing (SP) to ensure subcarriers of comparable quality are paired during the source-to-relay and relay-to-destination transmission time slots, and joint source-relay power allocation (PA) to subcarriers. It is demonstrated that the shorter the distance between the source and the relay, the more impact of multiuser diversity on the system capacity. By employing SA and joint source-relay PA only, a near-optimal performance can be achieved.

In the third contribution, an optimal asymmetric resource allocation (ARA) scheme is proposed for the asymmetric scenario. Unlike the symmetric resource allocation (SRA) scheme in the second contribution, the proposed ARA scheme does not require the source-to-relay and relay-to-destination transmission slots to be equal, and does not require bits to be transmitted in groups over paired subcarriers in the two consecutive slots. As a result, the overhead due to feedback information is reduced dramatically. Simulation results show that ARA achieves a higher system capacity, and has less sensitivity to channel estimation errors than existing methods, especially when there are a larger number of users, a longer source-to-destination distance, and when there is a significant difference between the source-to-relay and the relay-to-destination distances. Furthermore, the proposed ARA scheme has a fast convergence speed.

Acknowledgement

This thesis would not have been finished without loads of support and help from the following people. Dr. Xu Zhu, my supervisor, has my deepest gratitude for teaching me so much about research skills. This work benefits from her patient guidance, constant encouragement and invaluable comments. I also appreciate concerns from Dr. Yi Huang for my research.

I am grateful to the ORSAS committee of UK and the ORSAS committee of the University of Liverpool for financing me throughout my Ph.D. period.

I would like to thank the University of Liverpool, as well as the Department of Electrical Engineering and Electronics, for providing outstanding training and facilities to research students.

Warm thanks are given to Professor Guangxi Zhu and Dr. Weimin Wu at Huazhong University of Science and Technology for shaping my interests in wireless communications.

I would also like to thank my colleagues in the Signal Processing Group: Dr. Jingbo Gao for inspiring discussions and sharing views, Dr. Ye Wu and Dr. Mohamed Musbah for their support and encouragement, Nazmat Surajudeen-Bakinde and Linhao Dong for creating a family-like atmosphere in the lab where I have been working. I am pleasure to share my good and tough times with you.

Finally, my gratitude is dedicated to my parents. I would had no chance to pursue my goals in my life without their support, patience and encouragement. This thesis is dedicated to them.

Contents

List of Figures

List of Tables

Abbreviations and Acronyms

3G	third generation
3GPP	3rd generation partnership project
4G	fourth generation
AF	amplify-and-forward
ARA	asymmetric resource allocation
AWGN	additive white Gaussian noise
BER	bit error rate
BS	base station
CDMA	code division multiple access
CIR	channel impulse response
CNR	channel-to-nosie ratio
CPI	cyclic prefix insertion
CPR	cyclic prefix removal
CSI	channel state information
CoMP	coordinated multipoint

DF	decode-and-forward
DL	downlink
DRR	deficit round-robin
EDF	earliest deadline first
EV-DO	evolution-data optimized
FDE	frequency domain equalisation
FDM	frequency division multiplexing
FDMA	frequency division multiple access
FFT	fast fourier transform
FIFO	first-in first-out
FQ	fair queueing
GA	genetic algorithm
HDR	high data rate
HOL	head-of-line
HSPA	high speed packet access
IFFT	inverse fast fourier transform
ISI	intersymbol interference
KKT	Karush-Kuhn-Tucker
LTE	long term evolution
MAC	medium access control

MC	maximum capacity
MDU	max-delay-utility
MIMO	multiple-input and multiple-output
MS	mobile station
MWSC	maximum weighted sum capacity
M-LWDF	modified largest weighted delay first
OFDM	orthogonal frequency-division multiplexing
OSI	open system interconnection
PA	power allocation
PD	packet dependent
PF	proportional fair
PHY	physical
PL	path loss
QoS	quality of service
RA	resource allocation
RMS	root-mean-square
RR	round-robin
RS	relay station
SA	subcarrier allocation
SC-FDMA	single-carrier FDMA

SNR	signal-to-noise ratio
SOFDMA	scalable OFDMA
SP	subcarrier pairing
SRA	symmetric resource allocation
TDD	time division duplexing
TDMA	time division multiple access
UL	uplink
UMTS	universal mobile telecommunications system
VBR	variable bit rate
WFQ	weighted fair queuing
WiMAX	worldwide interoperability for microwave access

Chapter 1

Introduction

1.1 Background

Wireless communications have experienced a rapid growth over the past decades [1, 2], and a surge of research activities have been carried out to provide a higher system capacity [3], a better quality of service (QoS) [4], lower power consumption [5], and a more flexible system coverage [6]. In order to satisfy these aforementioned demands, various technologies such as orthogonal frequency-division multiplexing (OFDM) [7] and relay have been proposed and applied to wireless communication systems.

The OFDM technique, which divides a broadband channel into a set of orthogonal narrowband subcarriers, is capable of combatting the severe channel conditions, *e.g.*, narrowband interference and frequency-selective fading due to multipath, without complex equalization filters [2]. Therefore, OFDM is one of the promising techniques in the 3rd Generation Partnership Project (3GPP) Long Term Evolution (LTE) and Worldwide Interoperability for Microwave Access (WiMAX) systems, which are pre-fourth generation (pre-4G) technologies, as well as LTE-Advanced and WiMAX 2, which are known as 4G technologies.

Adaptive resource allocation (RA) plays a crucial role in OFDM based wireless

systems [8, 9]. First of all, different subcarriers can be allocated to different users to provide a flexible multiuser access scheme [10]. Since channel responses vary on different subcarriers for different users, *i.e.*, subcarriers suffering deep fading for one user may be in good condition for another, subcarrier allocation enhances the system performance by introducing multiuser diversity. Secondly, powers on different subcarriers can differ from each other with the OFDM system, which also increases the degree of freedom of the system, and results in a better system performance. Hence, the OFDM based system allows a wide range of adaptive RA schemes to be applied, which improves the performance of wireless networks.

Conventional communication networks, where each layer is designed to operate independently, do not utilize the spectrum and energy efficiently, especially in wireless networks which are confronted with time-varying fading channels, limited bandwidth, and competition of air resources among multiple users. Furthermore, heterogenous traffic in wireless networks and mobile terminals requires seamless and adaptive QoS [4], which includes not only the system capacity and bit error rate (BER) at the physical (PHY) layer, but also the delay tolerant, retransmission number, outage probability and so on at upper layers. Therefore, a cross-layer design of the wireless network is desirable.

Cooperative relaying was first proposed in the 1970s [11] and has drawn much attention recently for its applications in wireless communications [12]. By employing relays between the base station and mobile terminals to provide enhanced coverage of mobile services, system capacity, and flexibility for network extension, cooperative relaying is the key technology for LTE-Advanced/WiMAX 2 systems.

This thesis first presents the cross-layer design by considering the system capacity, delay tolerance, data size and QoS priority for LTE/WiMAX systems. The cross-layer design consists of an RA scheme at the PHY layer and a scheduling scheme at the medium access control (MAC) layer. After that RA for the relay based LTE-Advanced/WiMAX 2 system is investigated, where the RA consists

of subcarrier allocation (SA), subcarrier pairing (SP) and power allocation (PA). Optimal RA scheme for asymmetric relay systems is then investigated, where the transmission durations are not compulsorily equal.

1.2 Research Contributions

The research conducted during this PhD study has produced the following main contributions:

- An adaptive cross-layer design with the packet dependent (PD) scheduling at the MAC layer, for LTE/WiMAX systems with heterogeneous traffic, is proposed. The proposed PD scheduling scheme provides significant performance advantages over the queue based scheduling schemes in terms of the system throughput, BE (best effort) traffic throughput, QoS traffic delays and QoS traffic outage probabilities, with a wide range of the number of users, and a moderate to high signal-to-noise ratio (SNR) range. It is also proven to achieve the maximum system stability region. By properly choosing the number of packets used for weight calculation, our user based cross-layer design requires a much lower overall complexity than the conventional queue based designs. The proposed suboptimal RA algorithm provides a performance close to that of the optimal algorithm, at a much lower complexity.

- RA for symmetric regenerative relay based LTE-Advanced/WiMAX 2 systems is investigated, and the optimal RA scheme for the system is proposed, which consists of SA, SP and joint source-relay PA. Among the three steps of the proposed optimal RA, SA has a dominant impact on the system capacity when the relay is closer to the source than to the destinations, due to multiuser diversity, and joint source-relay PA has a larger impact on the system capacity than SP. A suboptimal RA scheme, which consists

of the proposed subcarrier allocation and joint source-relay PA algorithms only, provides a near-optimal performance. Furthermore, this work can be extended to the multi-hop multi-destination OFDM relay system.

- The optimal asymmetric RA scheme is proposed for the regenerative relay based LTE-Advanced/WiMAX 2 system, by relaxing the constraints that two subcarriers in consecutive time slots should be paired for data transmission, and durations in consecutive time slots should be equal for data transmission. The proposed RA scheme increases the degree of freedom for transmission. Compared to existing RA schemes, the optimal asymmetric RA scheme requires much less feedback information, and achieves a higher system capacity especially when there is a significant difference between the distance from the source to the relay and the distance from the relay to the destination. The optimal asymmetric RA scheme also demonstrates less sensitivity to channel estimation errors, and a fast convergence speed.

1.3 Thesis Organisation

The rest of this thesis is organised as follows. The wireless communication channels and systems are introduced in Chapter 2. Chapter 3 presents the literature survey on resource management of broadband wireless communication systems. A cross-layer design for LTE/WiMAX systems is proposed and discussed in Chapter 4. RA for the relay based LTE-Advanced/WiMAX 2 system is investigated in Chapter 5, where both the symmetric and asymmetric time duration scenarios are considered. Conclusions and future work are presented in the final chapter.

1.4 Publication List

A number of publications during the study of this research are listed below, which all contribute to the thesis.

Journal Papers

1. N. Zhou, X. Zhu, J. B. Gao and Y. Huang, "Optimal asymmetric resource allocation with limited feedback for OFDM based relay systems," *IEEE Trans. Wireless Commun.*, vol. 9, pp. 552-557, Feb. 2010.

2. N. Zhou, X. Zhu, Y. Huang and H. Lin, "Low complexity cross-layer design with packet dependent scheduling for heterogeneous traffic in multiuser OFDM systems," *IEEE Trans. Wireless Commun.*, vol. 9, pp. 1912-1923, Jun. 2010.

3. N. Zhou, X. Zhu, Y. Huang and H. Lin, "Optimal resource allocation for OFDM based multi-destination relay systems," submitted to *IET Commun.*

4. N. Zhou, X. Zhu and Y. Huang, "Optimal asymmetric resource allocation and analysis for OFDM based multi-destination relay systems in the downlink," submitted to *IEEE Trans. Veh. Technol.*

Conference Papers

1. N. Zhou, X. Zhu, and Y. Huang, "Cross-layer optimization with guaranteed QoS for wireless multiuser OFDM systems," in *Proc. IEEE PIMRC'2007*, Athens, Greece, Sep. 2007.

2. N. Zhou, X. Zhu, Y. Huang, and H. Lin, "Novel batch dependent cross-layer scheduling for multiuser OFDM systems," in *Proc. IEEE ICC'2008*, pp. 3878-3882, Beijing, China, May 2008.

3. N. Zhou, X. Zhu, Y. Huang, and H. Lin, "Adaptive resource allocation for multi-destination relay systems based on OFDM modulation," in *Proc. IEEE ICC'2009*, Dresden, Germany, Jun. 2009.

4. N. Zhou, X. Zhu, J. B. Gao, and Y. Huang, "Dynamic resource allocation with limited feedback for OFDM based cooperative networks," in *Proc. IEEE ICC'2009*, Dresden, Germany, Jun. 2009.

5. N. Zhou, X. Zhu, and Y. Huang, "Genetic algorithm based cross-layer resource allocation for wireless OFDM networks with heterogeneous traffic," in *Proc. EUSIPCO'2009*, Glasgow, United Kingdom, Aug. 2009.

6. N. Zhou, X. Zhu, and Y. Huang, "Optimal asymmetric resource allocation for OFDM based multi-destination relay systems in the downlink," submitted to *Proc. IEEE ICC'2011*, Kyoto, Japan, Jun. 2011.

7. L. H. Dong, X. Zhu, N. Zhou and Y. Huang, "Fairness based asymmetric resource allocation for decode-and-forward multi-relay systems," submitted to *Proc. IEEE ICC'2011*, Kyoto, Japan, Jun. 2011.

Chapter 2

Wireless Communication Channels and Systems

This chapter describes wireless communication channels, and reviews various wireless communication systems. The fundamentals of OFDM and relay techniques are also presented.

2.1 Wireless Communication Channels

2.1.1 Large-Scale Path Loss

The signal strength reduction caused by large distances between the transmitter and the receiver is referred to as large-scale path loss, simplified as path loss (PL). After being transmitted, signals decay as a function of d, the distance between the transmitter and the receiver, and the received signal power $P_r(d)$ can be expressed as

$$P_r(d) = \frac{P_t G_t G_r \lambda^2}{(4\pi)^2 d^\tau L} \tag{2.1}$$

where P_t is the transmitted power, G_t and G_r are the transmitter and receiver antenna gains, respectively, L indicates the system path loss factor, τ is the PL

Table 2.1: PL exponent values in different environments

Environment	PL exponent value
In building line-of-sight	1.6 to 1.8
Free space	2
Obstructed in factories	2 to 3
Urban area cellular	2.7 to 3.5
Shadowed urban cellular	3 to 5
Obstructed in building	4 to 6

exponent, and λ is the wavelength. The value of τ is normally in the range of 1.6 to 6 as shown in Table 2.1.

The PL, $PL(d)$, is defined as the difference between the transmitted power and the received power in dB, and it is expressed by

$$PL(d) = 10\log\left(\frac{(4\pi)^2 d^\tau L}{G_t G_r \lambda^2}\right) \tag{2.2}$$

It can be seen from (2.2) that PL can also be written as

$$PL(d) = C + 10\tau \log d \tag{2.3}$$

where $C = 10\log\left(\frac{(4\pi)^2 L}{G_t G_r \lambda^2}\right)$. PL is a major component in analysis and planning of a wireless communication system.

2.1.2 Small-Scale Fading

Small-scale fading, refers to the rapid fluctuation of channel gains and change of phases over a short time or distance, where the large-scale PL can be ignored [13]. Fading is usually caused by multipath, which is common in urban areas. In urban environments, transmitters and receivers are surrounded by structures, trees, moving vehicles and pedestrians, which cause reflection, diffraction and scattering of the transmitted signals. Therefore, the signal undergoes different paths, and each of the paths generates a unique wave of the transmitted signal with

the randomly distributed amplitude, phase and delay. The receiver vectorially combines all these waves, and this causes the signal distortion and fading.

2.1.2.1 Factors and Parameters of Small-Scale Fading

Letting f_i and σ_i denote the channel gain and delay for the ith path, respectively, the root-mean-square (RMS) delay spread σ is expressed by

$$\sigma = \sqrt{\frac{\sum_i f_i^2 \sigma_i^2}{\sum_i f_i^2} - \left(\frac{\sum_i f_i^2 \sigma_i}{\sum_i f_i^2}\right)^2} \tag{2.4}$$

The RMS delay spread interprets the multipath richness. In other words, the higher the RMS delay spread is, the larger the effect of multipath is.

In wireless communications, coherence bandwidth B_c is defined as the range of frequencies where two frequency components have a strong potential for amplitude correlation. Coherence bandwidth B_c can be derived from the RMS delay spread σ, *i.e.*, $B_c \approx \frac{1}{5\sigma}$.

The relative motion between the base station and the mobile user causes Doppler shift f_d, which is given by

$$f_d = \frac{v}{\lambda} \cos\theta \tag{2.5}$$

where v represents the velocity of a mobile user moving at, and θ is the angle between the direction of the received signal wave and the direction of the mobile user's motion. $f_m = \frac{v}{\lambda}$ denotes the maximum Doppler shift. Doppler spread B_d denotes the difference between Doppler shifts of signal components, and $B_d = f_m$ in most cases.

The coherence time T_c is referred to as the time over which two received signal have a strong potential for amplitude correlation [13]. It can be expressed in terms of the Doppler shift f_m, *i.e.*, and $T_c \approx \frac{9}{16\pi f_m}$.

Table 2.2: Small-scale fading effects

Flat Fading	Signal Bandwidth < Coherence Bandwidth
Frequency Selective Fading	Signal Bandwidth > Coherence Bandwidth
Fast Fading	Coherence Time < Symbol Period
Slow Fading	Coherence Time ≫ Symbol Period

2.1.2.2 Types of Small-Scale Fading

Depending on the characteristics of the transmitted signal, *i.e.*, singal bandwidth and symbol period, and the nature of the multipath channel, *i.e.*, RMS delay spread σ and Doppler spread B_{d}, there are four different fading effects, which are flat fading and frequency selective fading due to multipath delay spread, and fast fading and slow fading due to Doppler spread. These fading effects are shown in Table 2.2.

- Flat Fading — If the RMS delay spread of a wireless channel is smaller than the symbol period ($\sigma < T_{\mathrm{s}}$), *i.e.*, the channel coherence bandwidth is greater than the bandwidth of transmitted signal ($B_{\mathrm{c}} > B_{\mathrm{s}}$), all ranges of frequency over the transmission bandwidth will have approximately the same channel gain and linear phase. In this case, the channel is referred to as flat fading. Flat fading is the most common type of fadings described in the literature. The characteristics of flat fading channel gains vary in time according to different distributions such as Nakagami fading, Rician fading and Rayleigh fading [13].

- Frequency Selective Fading — If the RMS delay of the mobile radio channel is larger than the symbol period ($\sigma > T_{\mathrm{s}}$), or the channel coherence bandwidth is smaller than the signal bandwidth ($B_{\mathrm{c}} < B_{\mathrm{s}}$) in other words, the waves with long delays of the last transmitted symbol will reach the receiver while some waves with short delays of the current symbol arriving at the receiver. Under this circumstance, previous transmit symbols cause interference to current transmit symbols, which is called inter-symbol inter-

ference (ISI). In the frequency domain, the ISI is presented by a formation that frequency components of the received signal spectrum undergo different amplitudes. Therefore, it is referred to as frequency selective fading.

- Fast Fading — In a fast fading channel, the channel changes so rapidly that the signal undergoes different channels within a symbol duration. In other words, the coherence time of the channel is smaller than the symbol duration, *i.e.*, $T_c < T_s$ and $B_s < B_d$.

- Slow Fading — The channel remaining unchanged over one or several symbol periods is called a slow fading channel, where coherence time is much larger than the symbol period for a slow fading channel, *i.e.*, $T_c \gg T_s$. It also means that the Doppler spread of the channel is mush less than the symbol bandwidth ($B_s \gg B_d$).

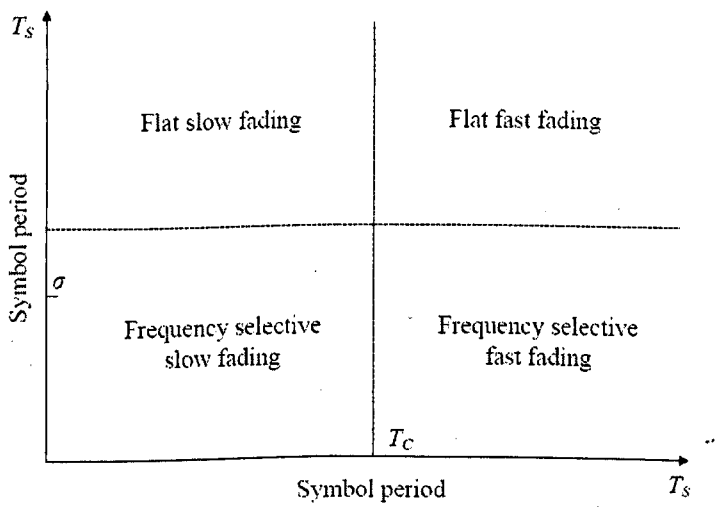

Figure 2.1: Fading types classified by the symbol period

According to the aforementioned fading effects, the small-scale fading can be classified into four types, flat slow fading channel, flat fast fading channel, frequency selective slow fading channel and frequency selective fast fading channel.

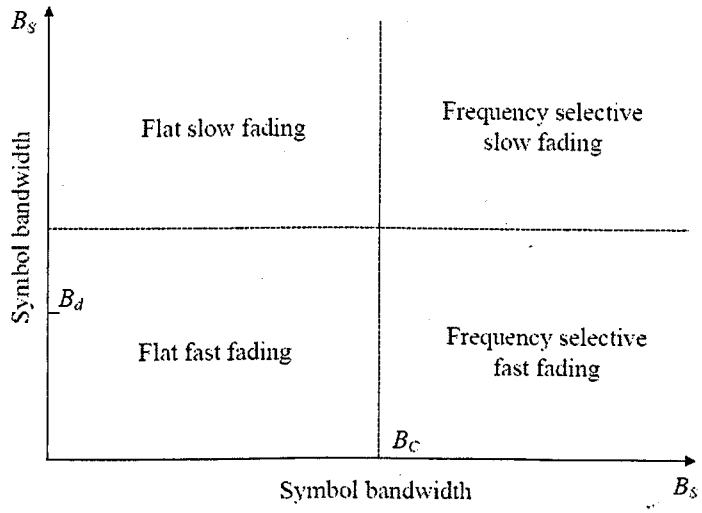

Figure 2.2: Fading types classified by the symbol bandwidth

These four channels are illustrated in Figures 2.1 and 2.2 as a function of symbol period and symbol bandwidth, respectively. This thesis places a focus on frequency selective slow fading channels.

2.1.2.3 Channel Model

The channel model is introduced here. It is assumed that each symbol is transmitted via signal bandwidth B_s during symbol period T_s. Supposing there are N_p paths between the transmitter and the receiver, the channel impulse response (CIR) is given by:

$$f(t) = \sum_{i=0}^{N_p-1} f_i \delta(t - \sigma_i) \tag{2.6}$$

where f_i and σ_i are the channel gain and delay for the ith path, respectively. Note that when $N_p = 1$, the channel reduces to flat fading.

Assuming that paths with different delays and different spatial diversity branches are independent, f_i is an independent zero mean complex Gaussian random variable, with a variance following the discrete exponential power delay profile:

$$E|f_i|^2 = b \cdot \exp(-\sigma_i/\sigma) \tag{2.7}$$

where b is the normalizing factor, and σ is the RMS delay spread. The Rayleigh distribution is commonly used to describe $|f_i|$, of which the probability density function is:

$$p(|f_i|) = \frac{|f_i|}{\tau^2} \exp\left(-\frac{|f_i|^2}{2\tau^2}\right) \qquad 0 \leq |f_i| < \infty \tag{2.8}$$

where τ is the RMS voltage value of the received signal, and τ^2 denotes the time-average power of the received signal. A typical Rayleigh fading channel CIR is illustrated in Figure 2.3.

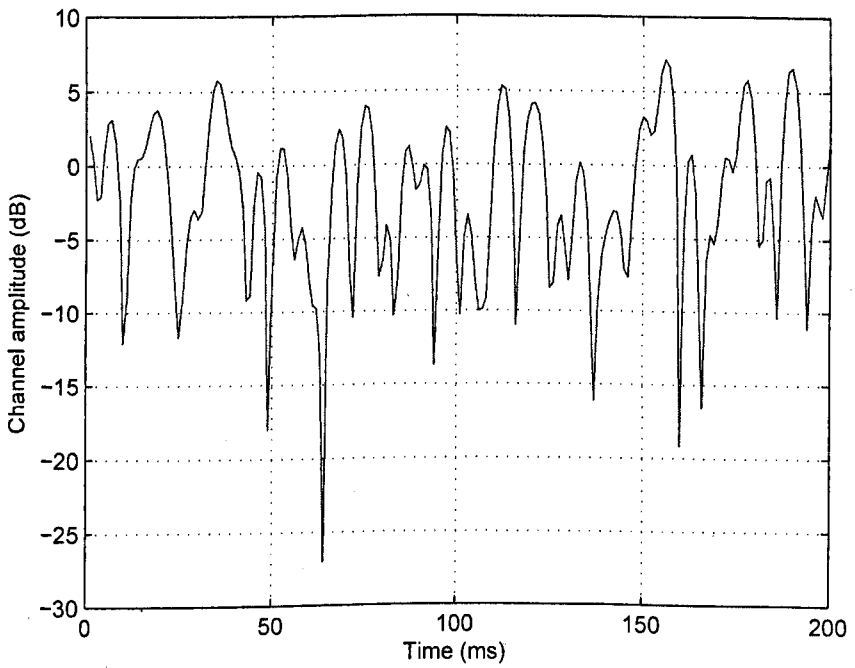

Figure 2.3: A typical Rayleigh fading channel impulse response

Letting $g(t)$ denote the pulse shape with the effects of the transmit filter and receive filter, the overall CIR is the convolution of the physical CIR $f(t)$ with $g(t)$:

$$h(t) = \sum_{i=0}^{N_p-1} f_i g(t - \sigma_i) \tag{2.9}$$

2.2 Wireless Communication Systems

2.2.1 Overview of Wireless Communication Systems

Since Guglielmo Marconi conducted the first wireless communication experiment in 1897, and the cellular systems were launched in 1980s, the means of communications between people have changed dramatically. People are more and more dependent on wireless communications, *e.g.*, the number of subscribers of cellular systems all over the world grew from 25,000 in 1984 to about 1 billion in 2001 [13]. Not only to provide voice services, mobile network operators have deployed the third generation (3G) systems from 2001 onwards, such as universal mobile telecommunications systems (UMTSs) and code division multiple access (CDMA) 2000 systems [14], to satisfy the rapidly increasing demand for various data requests.

Table 2.3: Evolution of mobile telephony standards

2G	3G	3.5G	pre 4G	4G
GSM	UMTS	HSPA	LTE	LTE-Advanced
cdmaOne	CDMA2000	EV-DO	UMB (withdrawn)	–
–	–	–	WiMAX (IEEE 802.16e)	WiMAX 2 (IEEE 802.16m)

From the beginning of the 21st century, demand for data services has been increasing dramatically. Hence, evolution-data optimized (EV-DO) [15] and high speed packet access (HSPA) [16], which are regarded as 3G transitional (3.5G) systems, were respectively operated in 2003 and 2005, to provide mobile broadbands over cellular networks. The maximum download speeds for HSPA and EV-DO are 14.4 Mbps and $4.9 \times N$ Mbps, respectively, and their maximum upload speeds are 5.76 Mbps and $1.9 \times N$ Mbps, respectively, where N is the number of 1.25 MHz spectrum chunks used in EV-DO systems.

The evolution of HSPA, which is referred to as HSPA+ [17], was proposed to

enhance the system capacity of HSPA. The downlink and uplink peak data rates of HSPA+ are $42 \times N$ and $11 \times N$, respectively, where N is the number of 5 MHz carriers employed. In recent years, pre-4G standards, WiMAX and LTE, have become commercial in North America and Europe, respectively. WiMAX, a novel technique based on the IEEE 802.16 standards, can provide broadband access with download and upload speeds up to 75 Mbps and 25 Mbps, respectively. On the other hand, LTE offers 326.4 Mbps and 86.4 Mbps as the peak downlink and uplink data rates. Based on WiMAX and LTE, WiMAX 2 and LTE-Advanced are proposed as the 4G telecommunication techniques, to support peak downlink data rate up to 1 Gbps and peak uplink data rate up to 500 Mbps [18]. All the aforementioned wireless communication techniques are summarized in Table 2.3 and Table 2.4.

2.2.2 OFDM Technique

By employing advanced radio techniques, the next generation wireless communication standards achieve a high spectral efficiency, *e.g.*, the downlink and uplink peak spectral efficiencies are respectively 30 bps/Hz and 15 bps/Hz in LTE-Advanced systems. Even though, the demand for high system capacity urges a wide spectrum for all the next generation wireless communication standards. For instance, LTE-Advanced provides an ultra high capacity of up to 1 Gbps over a wide spectrum up to 100 MHz. However, the higher the data rate, the larger the impact of multipath fading. OFDM [7] is one of the effective solutions to frequency selective fading. Compared to previous techniques, OFDM requires less complex equalisation filters, and it has been selected as one of the key techniques in pre-4G and 4G standards.

The procedure of the OFDM system is presented in Figure 2.4. Suppose there are N symbols $(S[0], ..., S[N-1])$ waiting for transmission. After the inverse fast

Table 2.4: Comparison of mobile telephony standards

Family	Standard	Channel Bandwidth (MHz)	Peak Data Rates (Mbps)	Radio Tech
3GPP2	CDMA2000	1.25	0.307 (DL) 0.307 (UL)	CDMA
	EV-DO Rev.B	$1.25 \times N^1$	$4.9 \times N^1$(DL) $1.9 \times N^1$(UL)	TDMA (DL) CDMA (UL)
	UMB	$1.25 \times N^1$	$18.7 \times N^1$(DL) $4.3 \times N^1$(DL)	OFDMA MIMO
3GPP	UMTS	5	2.0 (DL) 2.0 (UL)	WCDMA
	HSPA	5	14.4 (DL) 5.76 (UL)	WCDMA
	HSPA+	$5 \times N^2$	$42 \times N^2$(DL) $11 \times N^2$(UL)	WCDMA MIMO
	LTE	Scalable up to 20	300 (DL) 75 (UL)	OFDMA (DL) SC-FDMA (UL) MIMO
	LTE-Advanced	Up to 100	1000 (DL) 500 (UL)	OFDMA (DL) SC-FDMA (UL) MIMO Relay CoMP
IEEE	WiMAX	5, 7, 8.75, 10	128 (DL) 66 (UL)	SOFDMA MIMO
	WiMAX 2	5 to 20	300 (DL) 135 (UL)	SOFDMA MIMO Relay

[1] N =number of carriers, max. 15
[2] N =number of carriers, max. 4

fourier transform (IFFT), the transmission data can be presented as:

$$x[n] = \sum_{i=0}^{N-1} S[i] e^{j\frac{2\pi}{N}in} \quad n \in \{0, ..., N-1\} \tag{2.10}$$

The received signals are

$$y[n] = h(n) \otimes x[n] + z[n] \quad n \in \{0, ..., N-1\} \tag{2.11}$$

where $h(n)$ and $z[n]$ are the CIR and noise. The FFT is applied to the received

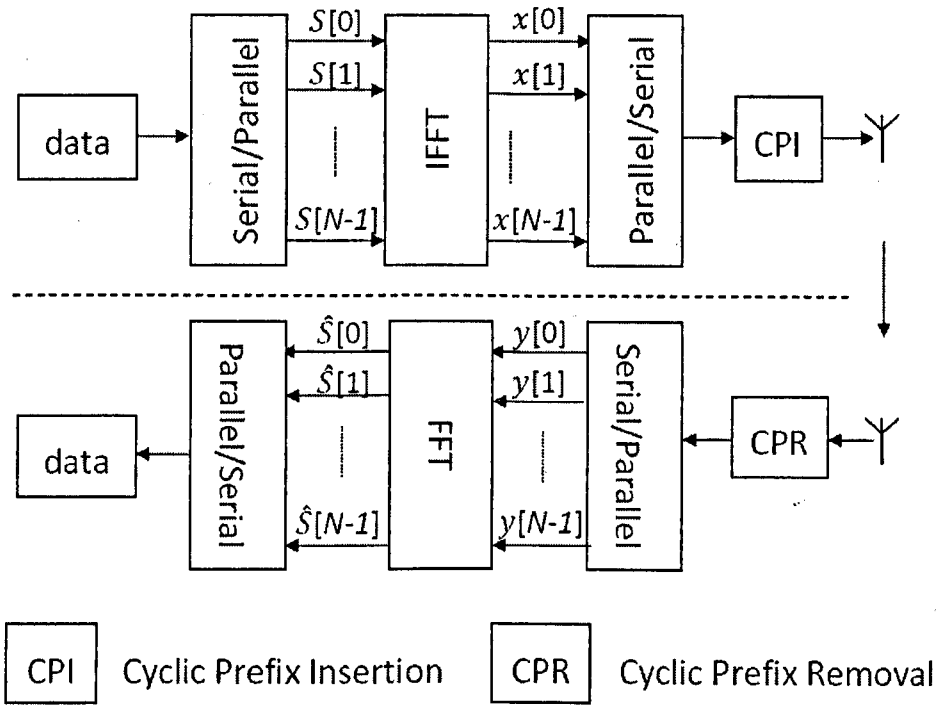

Figure 2.4: Block diagram of the OFDM transmission system

signals after CPR, and recovered signals are

$$\begin{aligned}\widehat{S}[n] &= \sum_{i=0}^{N-1} y[i]e^{-j\frac{2\pi}{N}in} \quad n \in \{0,...,N-1\} \\ &= \text{FFT}\left(h(n) \otimes x[n] + z[n]\right) \\ &= H_n S[n] + Z[n] \end{aligned} \tag{2.12}$$

where H_n and $Z[n]$ are the CIR and noise on the nth subcarrier in the frequency domain.

It can be derived from (2.10) that each transmission data $x[n]$ consists of information from N symbols, and the information of one symbol has been spread across N transmission data. Hence, the symbol period in OFDM systems becomes NT_s, leading to a mitigated ISI. In the frequency domain, the frequency range for each symbol is narrowed to B_s/N so that each transmission data carried on one subcarrier undergoes flat fading.

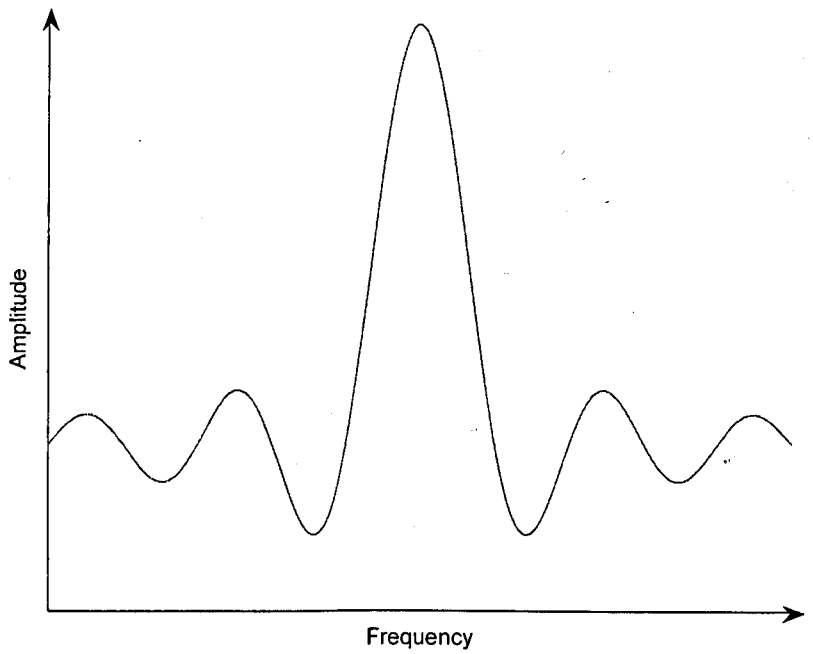

Figure 2.5: Waveform of one OFDM subcarrier

There are several benefits brought by the OFDM technique. First of all, in OFDM systems, each subcarrier is in flat fading, and therefore the ISI problem in multipath environments is eliminated. Secondly, since the bandwidth of each subcarrier is narrowed, the system is unlikely to suffer fading over all subcarriers in a multipath channel. Hence, there is frequency diversity in OFDM systems. Additionally, each subcarrier can be modulated independently in OFDM systems. It leads to a higher spectral efficiency than conventional single-carrier systems under frequency selective circumstances. Furthermore, OFDM leads to a higher efficiency in spectrum than conventional frequency division multiplexing (FDM), where the frequency bands (carriers) are not overlapped, and there is a guard band between two adjacent carriers. In OFDM systems, all the subcarriers are overlapped as shown in Figures 2.5 and 2.6, where one subcarrier achieves its peak attenuation when others are zeros. This is because of the help of IFFT. Therefore, OFDM achieves a higher spectral efficiency than FDM due to the

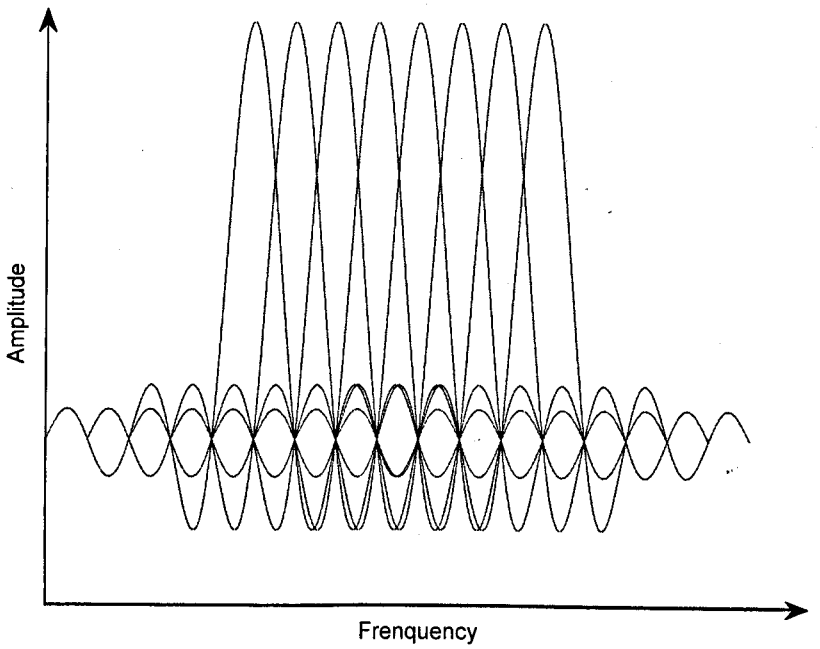

Figure 2.6: Waveform of eight OFDM subcarriers

removal of the guard bands.

2.2.3 Relay Technique

Since first proposed in the 1970s by Meulen [11], cooperative relaying has drawn much attention [12]. The basic concept of relaying is to employ relay nodes to help to improve the communication quality between the source node and the destination node.

The next generation wireless communication system normally utilizes a higher frequency than the carrier frequency used now due to the frequency availability. This results in more severe PLs. On the other hand, base stations in the future are to provide a larger coverage area than today's technology. Therefore, cooperative relaying is a promising technology for LTE-Advanced and WiMAX 2 systems, since it provides ubiquitous wide area coverage, flexible extension and capacity increase over traditional wireless networks [19].

Depending on with or without a direct path between the source and the destination, relay systems can be categorized into two types. For the relay systems with a direct path, signals are firstly transmitted from the source to the relay and to the destination. After receiving the data, relay forwards them to the destination. Finally, the destination will combine all the signals it received from the source and the relay. This process can be regarded as the virtual multiple-input multiple-output (MIMO) technique. In the other type where the destination cannot hear the source, it only acquires the signal from the relay. This thesis considers the latter type of relay systems.

Based on how the relay node performs after it receives the signal from the source, there are two main relaying strategies: nonregenerative relaying and regenerative relaying [12]. In nonregenerative relaying systems, the relay simply amplifies the received signal and retransmits it to the destination. Hence, it is also called amplify-and-forward (AF) relaying. While in regenerative relaying systems, the relay decodes the signal and then forwards it to the destination. Therefore, regenerative relaying is also referred to as decode-and-forward (DF) relaying.

In [20], it is demonstrated that AF outperforms DF in terms of the system capacity when SNR between the source and the relay is low (under 0 dB), since the relay may not be able to decode the signal from the source. While at a moderate to high SNRs (> 5 dB) between the source and the relay, the DF mode provides higher system capacity than AF. This is because that AF could amplify the noise between the source and the relay while retransmitting the signal. In this thesis, DF relay systems are considered.

Chapter 3

Overview of Resource Management for Broadband Wireless Systems

This chapter presents a literature review on resource management for broadband wireless communication systems. In Section 3.1, diversity in the OFDM based wireless systems is described. Literature reviews on RA, scheduling and cross-layer designs for the OFDM based wireless system are presented in Sections 3.2, 3.3 and 3.4, respectively.

3.1 Diversity in OFDM based Broadband Wireless Systems

Resource management for OFDM based broadband wireless systems utilizes different kinds of diversity to provide better system performance. This section presents the most common diversity in OFDM systems, frequency diversity and multiuser diversity.

3.1.1 Frequency Diversity

As mentioned in Section 2.1, broadband wireless systems employ a wide spectrum in order to provide high capacity, and hence they suffer from the frequency selective fading. In this case, some parts of the spectrum are in deep fading while the channel quality is good for some other parts. Hence, OFDM technique, which divides the whole signal frequency band into many narrowbands with flat fading, produces frequency diversity out of frequency selective fading.

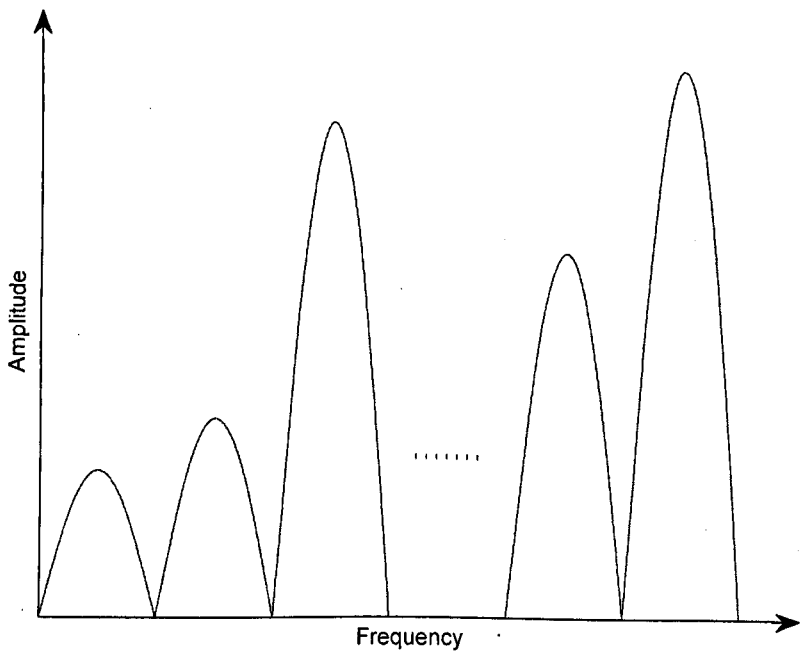

Figure 3.1: Different amplitude values of subcarriers

Different frequency bands have different properties. The exploration of the frequency diversity leads to a high degree of freedom for the transmission, resulting in a better performance. For example, employing the suitable modulation type in each frequency band helps with increasing the system capacity and decreasing the BER. In [21], the variable rate and power MQAM modulation scheme was proposed for a single-user wireless communication system. The RA for single-user OFDM systems was investigated in [22].

3.1.2 Multiuser Diversity

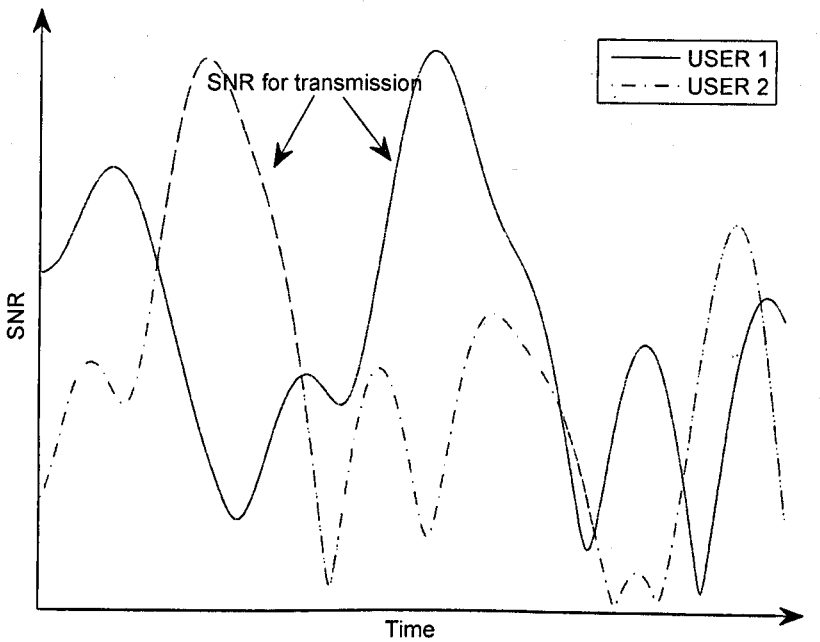

Figure 3.2: Scheduling for a two-user case

Time variation, referring to the fading changes over time, is the nature of wireless communication channels [1, 13]. Hence, the performance of point-to-point systems may vary in a wide range with times [23].

In multiuser wireless communication systems, the channels are independent across different users. For example, while the channel for user 1 is in fading, the channel may be in a good condition for the other user, as demonstrated in Figure 3.2. In this case, the system can always choose the user with a higher SNR to transmit, *i.e.*, $SNR = max\{SNR(1), SNR(2)\}$, leading to a higher system capacity than conventional systems. Hence, the flexibility of scheduling users' data in the multiuser wireless communication system increases the system capacity, and it is called multiuser diversity. Intuitively, when the number of users increases, the SNR for transmission ($max\{SNR(1), SNR(2), ...\}$) also augments, since there is more degree of freedom due to scheduling of different users. Hence,

the system capacity, or spectral efficiency in other words, benefits more from the multiuser diversity with the increase of the number of users.

Recently, much research has been carried out on multiuser diversity [3, 5, 10, 24, 25, 26, 27]. Scheduling has been investigated in [3] and [24] for uplink and downlink, respectively. It has been proved in [3] that, in order to maximize the downlink system capacity, the user with the best channel gain should be selected to transmit. A similar conclusion can be found in [24] for the uplink multiuser wireless communication system (multi-point-to-point). By utilizing multiuser diversity, fast scheduling was adopted for high-speed packet data services in the downlink [28], such as Qualcomm's High Data Rate (HDR) system [29] and the HSPA system [30].

3.2 PHY Layer Resource Allocation

In this section, the RA techniques at the PHY layer of OFDM systems are reviewed. Compared to the conventional single-carrier system, an OFDM system usually contains a very large number of subcarriers, which enable a higher degree of freedom for RA. In frequency selective fading channels, some subcarriers may be in deep fading. However, it is unlikely that all users experience deep fading on the same subcarrier, since the channel characteristics for different users are mutually independent. Hence, by assigning each subcarrier to the user with the best channel-to-nosie ratio (CNR) value on that subcarrier, a system capacity higher than the capacity achieved by single-carrier systems can be obtained, benefiting from frequency diversity and multiuser diversity.

In broadband wireless communication systems, since the system bandwidth is larger than the channel coherence bandwidth, different frequency bands usually hold different channel characteristics, such as log-normal shadowing, short term Rayleigh fading and noise [31]. Hence, uniform power allocation for the whole

bandwidth will cause degradations of the system spectral efficiency and the BER. When OFDM is employed, power on each subcarrier can vary [32, 33], since each subcarrier is independent from the others. The adaptive power [34] on each subcarrier enhances the degree of flexibility, and provides better bit rate and robustness of data transmission [35].

Much research has been carried out on SA and PA at the PHY layer of OFDM systems [3, 5, 36, 37, 38]. In [3], the optimal SA and PA schemes were proposed for OFDM systems. By allocating each subcarrier to the user with the best channel quality on it, and allocating power according to the water-filling algorithm [39], the maximum system capacity can be achieved [3]. Adaptive SA and PA schemes were proposed in [5], where the system power is minimized while achieving the target data rates. In [36], SA and PA were studied to maximize the minimum data rate among all the users for a multiuser OFDM system. The low-complexity SA and PA schemes were proposed in [37], which minimize the system power while providing desired BERs and data rates for different users. Shen et al [38] proposed suboptimal SA and PA algorithms to guarantee the users' proportional data rates while maximizing the system capacity.

RA also plays a crucial role in cooperative relay systems, which have been proposed for LTE-Advanced and WiMAX 2. Since OFDM is one of the key techniques in LTE-Advanced and WiMAX 2 systems, and allows a wide range of adaptive RA schemes to be applied to improve the QoS of wireless communication systems [8][40]. Most work on adaptive RA in relay systems has assumed OFDM modulation [41, 42, 43, 44, 6, 45].

In [41, 42, 43, 44], adaptive RA to maximize the system capacity for a three-node configuration of the cooperative network, where a single source communicates with a single destination through a relay, was investigated in [41, 42, 43, 44]. An SP algorithm was proposed in [41] for the AF relay system, where the bits received by the relay over a subcarrier in the first time slot are forwarded to

the destination over a specific subcarrier in the second time slot. In [42], PA is performed at the source and relay separately, and a suboptimal joint source-relay PA scheme for the AF relay systems was proposed. In [43], the optimal SP and joint source-relay PA schemes were proposed for DF relay systems.

Adaptive RA for the relay system with multiple relays and multiple destinations has also been studied. The work in [43] was extended to the multi-hop regenerative relay system in [44], and was then extended to the multi-destination regenerative relay system in [46]. In [45], the relay-node selection and SA were investigated for a single-source, multi-relay and multi-destination system. In [6], the utility maximization framework was adopted to investigate how to allocate resources, and select relay-node and relay-strategy jointly. In [47], the Nash bargaining solution was applied to allocate subcarriers in the relay-aided cellular network. In [48], an iterative algorithm was proposed for SA and PA in a multi-source multi-destination relay system, assuming that the relay utilizes the same subcarrier in the second slot as the source transmitted with in the first slot. In [49], the authors proposed an optimal RA scheme for the multi-destination relay system. Asymmetric time allocation was proposed in [50], which is however for a single destination relay system and lacks analysis of time allocation.

3.3 MAC Layer Scheduling

Since the 1990s, communication networks have started to support multimedia services [51]. For example, voice and video streamings, which require different delays and jitters, are delivered via the same network. It is vital and challenging to meet different requirements of different services Scheduling [52][53], which is performed at the MAC layer to determine which data should be sent out at what time, is a good solution to the problem.

First-in first-out (FIFO) [54] is a straightforward scheduling scheme in com-

munication networks, which is used in bridges, switches and routers deployed in communication networks to schedule data in route to their next destination [55]. The fair queueing (FQ) concept was originally proposed in [56]. FQ divides the network capacity into fractions equally, and each of the active queue consumes one fraction. Hence, each active queue is assigned the same capacity, and the FQ scheduling scheme achieves the max-min fairness [57, 58]. In [59], the weighted fair queueing (WFQ) scheduling was proposed. It is different from FQ that the amount of system capacity allocated to queues by WFQ are proportional to the weights of queues. Therefore, FQ can be seen as a special case of WFQ with identical weights for all the queues. WFQ was shown to guarantee the end-to-end delays as well as the fairness fair among traffic [60].

The round-robin (RR) scheduling [61] is another simple scheduling scheme, which gives opportunities for different services to transmit data. RR is regarded as one of the simplest scheduling schemes, and it provides close max-min fair throughputs in a network [62]. In [63], a modified RR scheme, referred to as deficit round-robin (DRR), was proposed. The idea of DRR is that the queues, which are not selected in the current round due to the overlength of data, will be compensated in next rounds [63].

The earliest deadline first (EDF) scheduling was proposed by Liu and Layland in 1973 [64], where a service with the earliest deadline will be transmitted first. The EDF scheduling has been proven optimal in terms of end-to-end delays in a single link network [65].

3.4 Cross-Layer Design

In the conventional open system interconnection (OSI) model, there are seven layers, and each layer is designed to operate independently, as shown in Figure 3.3. As a result, the OSI model does not utilize the spectrum and energy efficiently,

especially in wireless networks which are confronted with time-varying fading channels, limited bandwidth, and competition of air resources among multiple users [66]. Therefore, a cross-layer design of the wireless network is desirable [67].

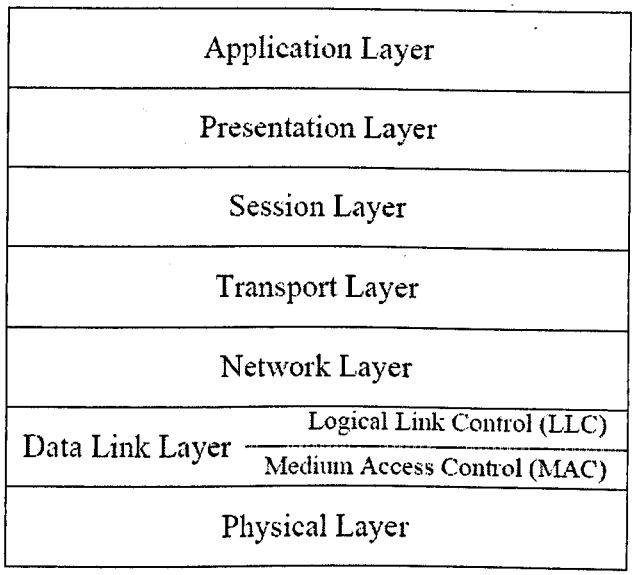

Figure 3.3: OSI model

Some work on RA has been conducted based on the cross-layer information [68, 69, 70]. Song et al [68, 69] used the utility function as a bridge between the PHY layer and the MAC layer. By optimizing the utility function, an optimal RA scheme was proposed to balance the spectral efficiency and data rate fairness. In [70], the ACK/NACK information from the MAC layer is used for RA at the PHY layer. The cross-layer RA algorithm achieves a higher spectral efficiency, compared to RA algorithms using unreliable channel state information (CSI).

Cross-layer scheduling has also been investigated [71, 72, 73, 74]. The proportional fair (PF) scheduling scheme was proposed by Qualcomm [71], which makes a trade-off between the fairness and the data rate. The PF scheduling scheme compares the ratios of instantaneous data rates to average data rate per user for difference users, and picks the user with the maximum ratio value to serve. With a small cost of system capacity, the PF scheduling scheme ensures the users with poor channel quality to obtain resources, implying the users with low SNR

will not starve, and a better fairness among users. The PF scheduling scheme is proven to be associated with the logarithmic utility function [75]. In [72], a modified largest weighted delay first (M-LWDF) scheduling scheme was proposed to satisfy delays and throughputs of all users by assigning a higher priority to the queues that have a larger head-of-line (HOL) packet delay relative to the delay bound, a higher instantaneous data rate relative to the average data rate, and a higher requirement for the outage probability. A cross-layer scheduling scheme based on the instantaneous data rates at the PHY layer and the delay bounds at the MAC layer was studied in [73]. An exponential scheduling scheme was studied in [74], which is optimal in terms of system throughput and provides a good QoS over a shared wireless link.

The cross-layer designs performing both the RA and scheduling have been investigated in [8, 23, 76]. In [8], a cross-layer design to satisfy the heterogeneous delay tolerances for services, was studied, where the delay tolerances are transformed into rate constraints. A max-delay-utility (MDU) cross-layer design was proposed in [23], which performs RA and scheduling simultaneously by maximizing a utility function of the delay. MDU provides a trade-off between the system capacity and the QoS requirements. In [76], the genetic algorithm (GA) was employed for RA and scheduling by maximizing the weighted sum capacity, where the weights are determined at the MAC layer according to the traffic delays.

Chapter 4

Resource Management for LTE/WiMAX Systems

With the increasing demands for high capacity and multi-tasking services, wireless communication systems are expected to support multiple users, each of whom has multiple heterogeneous traffic queues simultaneously. For example, a user can use a mobile device to make a phone call and download files from a website at the same time. Adaptive resource management plays a crucial role in multiuser multi-tasking services. Therefore, as aforementioned, the cross-layer design of the wireless network is desirable to achieve different QoS requirements simultaneously.

Most previous work on scheduling was traffic queue based [9, 23, 38, 72], where all packets contained in the selected queues are served until the data or the PHY resources are exhausted in each slot. This however leads to inefficiency if some packets in the unselected queues are more urgent than some packets in the selected queues. In [77], a packet based cross-layer scheduling approach was proposed for GSM/EDGE systems, which takes the packet delay and channel quality into account. However, it did not consider the system capacity. Furthermore, most previous work assumed that each user obtains a single traffic queue and did not consider a system with heterogeneous traffic for each user.

This chapter proposes a novel adaptive cross-layer design to maximize the weighted sum capacity of the LTE/WiMAX systems. This work is different in the following aspects. First, an OFDM system where each user has multiple heterogeneous traffic queues at the same time is considered. Second, a packet dependent scheduling scheme is proposed, which determines the packet transmission order by assigning different weights to different packets, and therefore is more efficient than the conventional queue dependent scheduling [72][23] where all packets in a queue have the same weight. The packet weights in PD scheduling are determined based on the delay, packet size and QoS priority level of the packets. Third, the proposed RA is user based, by maximizing the weighted sum capacity of users rather than queues, and therefore requires a lower complexity than the conventional queue based RA. The weight for each user is obtained by summing the weights of selected packets for that user. By properly choosing the number of selected packets for weight calculation, our proposed cross-layer design achieves a significant reduction of the overall complexity compared to the previous work [72][23]. Furthermore, an intensive performance analysis of the proposed cross-layer design is provided, which is proven to achieve the maximum system stability region. Simulation results show that, compared to the queue based M-LWDF [72] and MDU [23] scheduling schemes, the proposed PD scheduling scheme demonstrates significant performance advantages. Also, the proposed suboptimal RA algorithm provides a performance close to that of the optimal algorithm, while requiring a lower complexity.

After the system model introduced in Section 4.1, the problem is formulated in Section 4.2. The optimal and suboptimal MWSC based RA are presented in 4.3. Following that, PD scheduling is proposed in Section 4.4. Performance analysis and complexity analysis are respectively provided in Sections 4.5 and 4.6. Section 4.7 shows simulation results, and Section 4.8 presents the summary.

4.1 System Model

This chapter considers a general downlink time division duplexing (TDD) OFDM system in LTE/WiMAX networks, where the base station can acquire the CSI through the uplink dedicated pilots from all mobile stations at the beginning of each time slot, and use it for RA and scheduling. Figure 4.1 illustrates the multiuser OFDM system block diagram, where the subcarrier and power controller at the PHY layer performs subcarrier and power allocation, and the traffic controller at the MAC layer performs data scheduling. With a cross-layer design, the QoS information obtained by the traffic controller is transferred to the subcarrier and power controller for RA, and the RA results are fed back to the traffic controller in the base station for scheduling of the data to be sent out in each slot.

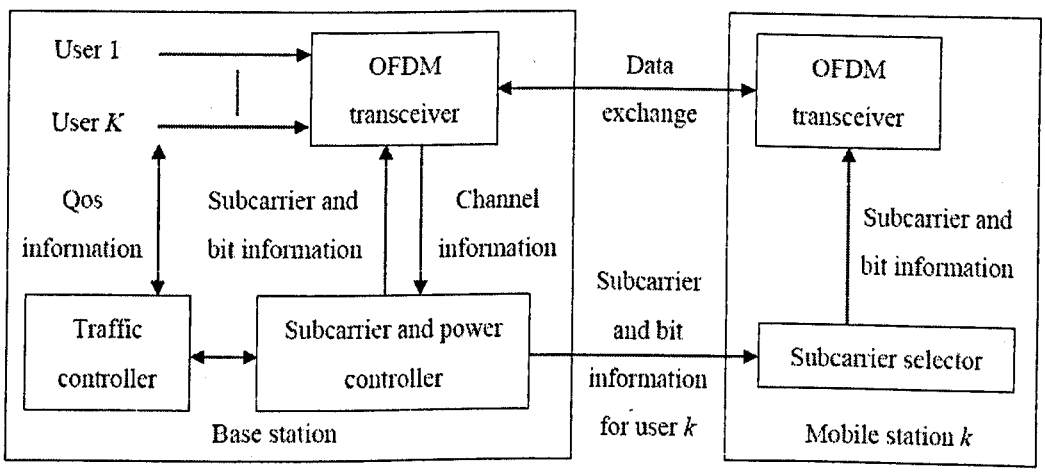

Figure 4.1: Multiuser OFDM system block diagram

It assumes a total bandwidth of B which is divided into N independent subcarriers and shared by K users. The total transmit power from the base station is assumed to be p_{T}. The OFDM signaling is time-slotted and each time slot is of length T_{slot} [9]. A quasi-static fading channel is considered, where the channel gain is constant during each slot and is independent of the channel for other slots. It is assumed that each user has I_k traffic queues with heterogeneous delay constraints for the queues. Define Ω_k as the index set of subcarriers allocated to user

k ($k \in \{1, ..., K\}$). To avoid the inter-user interference on the same spectrum, it is assumed that each subcarrier is allocated to only one user [3][38]. Let $p_{k,n}$ denote the power allocated to user k on subcarrier n ($n \in \Omega_k$), $h_{k,n}$ the corresponding channel gain, and N_0 the single-sided power spectral density of additive white Gaussian noise (AWGN). Assuming perfect channel estimation. The achievable instantaneous data rate of user k on subcarrier n is expressed as:

$$R_{k,n} = \frac{B}{N} \log_2(1 + p_{k,n}\gamma_{k,n}) \tag{4.1}$$

where

$$\gamma_{k,n} = \frac{|h_{k,n}|^2}{N_0 B/N} \tag{4.2}$$

is the channel-to-noise power ratio for user k on subcarrier n. The total achievable instantaneous data rate of user k is given by:

$$R_k = \sum_{n \in \Omega_k} R_{k,n} \tag{4.3}$$

4.2 Problem Formulation

A general maximum weighted sum capacity (MWSC) based cross-layer design, which is to maximize the weighted sum of all users' instantaneous capacities [51][76], is employed. Letting W_k denote the weight for user k, which contains the QoS information and is determined by scheduling at the MAC layer, the cost function to be maximized is given by:

$$J = \sum_{k=1}^{K} W_k R_k \tag{4.4}$$

subject to (C1) $p_{k,n} \geq 0$, (C2) $\sum_{k=1}^{K} \sum_{n \in \Omega_k} p_{k,n} \leq p_T$, (C3) $R_k T_{\text{slot}} \leq Q_k$, (C4) $\Omega_1 \cup \ldots \cup \Omega_K \subseteq \{1, 2, \ldots, N\}$, and (C5) $\Omega_k \cap \Omega_j = \emptyset$ ($k \neq j$), where Q_k is the

queue length of user k. The constraint in (C3) is to guarantee that no more resource is allocated to user k if the user already obtains sufficient resources to send all data out in current slot, to avoid waste of resources.

The MWSC based cross-layer design is general in that the weights W_k can be obtained by using a wide range of scheduling schemes. The RA scheme based on maximizing the sum capacity [3] can be regarded as a special case of the MWSC based cross-layer design with equal weights. In this thesis, a multi-objective optimization problem is investigated by taking into account not only the system capacity, but also the QoS traffic delay, QoS priority level and data amount. It is worth noting that there is generally not a single optimal solution to a multi-objective optimization problem. By projecting the information on delay, QoS priority level and data amount into the weights, the multi-objective optimization problem is formulated as a single cost function of MWSC as shown in (4.4), and it is able to obtain a unique optimal solution.

The above user based cross-layer design can be easily extended to the queue based cross-layer design, by replacing the user index k with the queue index i. Assuming that all users have ω queues each, *i.e.*, $I_k = \omega$ $(\forall k \in \{1, ..., K\})$, there are a total number of ωK queues in the system. The queue based MWSC cross-layer design [72][23] is to maximize the following cost function:

$$J = \sum_{i=1}^{\omega K} W_i R_i \tag{4.5}$$

subject to (C6) $p_{i,n} \geq 0$, (C7) $\sum_{i=1} \sum_{n \in \Omega_i} p_{i,n} \leq p_{\mathrm{T}}$, (C8) $R_i T_{\mathrm{slot}} \leq Q_i$, (C9) $\Omega_1 \cup \ldots \cup \Omega_{\omega K} \subseteq \{1, 2, \ldots, N\}$, and (C10) $\Omega_i \cap \Omega_j = \emptyset$ $(i \neq j)$, where W_i, R_i and Q_i denote the weight, instantaneous data rate and queue length of queue i, and Ω_i is the index set of subcarriers allocated to queue i.

Since the number of users is usually smaller than the number of queues (when $\omega > 1$), with given weights, the user based MWSC cross-layer design yields a

lower complexity than the conventional queue based cross-layer design.

4.3 Maximum Weighted Sum Capacity based Resource Allocation

In this section, an optimal algorithm and a suboptimal algorithm for RA based on the MWSC criterion is presented.

4.3.1 Optimal MWSC based Resource Allocation

This subsection presents the optimal MWSC based RA of (4.4). The optimal RA of (4.5) can be obtained by using similar method.

The problems in (4.4) involves both the integers and continuous variables Ω_k and $p_{k,n}$, and it is referred to as a mixed combinatorial and convex problem, which is NP-hard. Therefore, the spectrum-sharing factor $s_{k,n} \in (0, 1]$ is introduced, which indicates how long subcarrier n can be occupied by user k, to make the optimization problems traceable. Therefore, the instantaneous data rate of user k on subcarrier n is changed to:

$$R_{k,n} = \frac{Bs_{k,n}}{N} \log_2 \left(1 + \frac{p_{k,n}\gamma_{k,n}}{s_{k,n}}\right) \tag{4.6}$$

It can be derived from (4.6) that when $s_{k,n}$ is close to zero, $R_{k,n}$ also approaches zero. And the optimization problem in (4.4) is rewritten as:

$$\max_{s_{k,n},p_{k,n}} \sum_{k=1}^{K} W_k \sum_{n=1}^{N} \frac{Bs_{k,n}}{N} \log_2 \left(1 + \frac{p_{k,n}\gamma_{k,n}}{s_{k,n}}\right) \tag{4.7}$$

subject to (C11) $p_{k,n} > 0$, (C12) $\sum_{k=1}^{K}\sum_{n=1}^{N} p_{k,n} \leq p_{\mathrm{T}}$, (C13) $R_k T_{\mathrm{slot}} \leq Q_k$, (C14) $s_{k,n} \in [0, 1]$, and (C15) $\sum_{k=1}^{K} s_{k,n} = 1$ $(\forall n \in \{1, ..., N\})$.

First of all, it can be proven that the optimization problem (4.7) is concave.

Since the Karush-Kuhn-Tucker (KKT) conditions are necessary and sufficient for a solution being optimal to a convex optimization problem, the optimal solution satisfys the KKT conditions.

Considering (4.6), the Hessian of $R_{k,n}$ ($\forall k \in \{1,...,K\}, \forall n \in \{1,...,N\}$) is given by:

$$\begin{aligned} H(R_{k,n}) &= \begin{pmatrix} \frac{\partial^2 R_{k,n}}{\partial s_{k,n}^2} & \frac{\partial^2 R_{k,n}}{\partial s_{k,n}\partial p_{k,n}} \\ \frac{\partial^2 R_{k,n}}{\partial p_{k,n}\partial s_{k,n}} & \frac{\partial^2 R_{k,n}}{\partial p_{k,n}^2} \end{pmatrix} \\ &= \frac{B p_{k,n}\gamma_{k,n}^2}{N\left(s_{k,n}+p_{k,n}\gamma_{k,n}\right)^2} \begin{pmatrix} -\frac{p_{k,n}}{s_{k,n}} & 1 \\ 1 & -\frac{s_{k,n}}{p_{k,n}} \end{pmatrix} \end{aligned} \tag{4.8}$$

Since B, N, $p_{k,n}$ and $s_{k,n}$ are positive, the Hessian of $R_{k,n}$ is negative semidefinite, meaning that $R_{k,n}$ is a concave function. Hence, (4.7) is also concave.

Secondly, it is find the solutions satisfying the KKT conditions. The Lagrangian of (4.7) is:

$$\begin{aligned} L &= \sum_{k=1}^{K} W_k \sum_{n=1}^{N} s_{k,n} R_{k,n} - \zeta\left(\sum_{k=1}^{K}\sum_{n=1}^{N} s_{k,n}p_{k,n} - p_{\mathrm{T}}\right) \\ &\quad -\mu_k\left(\sum_{n=1}^{N} s_{k,n}R_{k,n}T_{\mathrm{slot}} - Q_k\right) - \sum_{n=1}^{N}\theta_n\left(\sum_{k=1}^{K} s_{k,n} - 1\right) \end{aligned} \tag{4.9}$$

where ζ, μ_k and θ_n are nonnegative Lagrange multipliers. It is straightforward that the Lagrangian (4.9) is concave, where its derivatives are decreasing functions. Letting ζ^*, μ_k^* and θ_n^* denote the optimal Lagrange multipliers, it can be obtained that:

$$\begin{aligned} \left.\frac{\partial L}{\partial p_{k,n}}\right|_{p_{k,n}=p_{k,n}^*, s_{k,n}=s_{k,n}^*} &= \frac{B\gamma_{k,n}}{N\left(1+p_{k,n}^*\gamma_{k,n}\right)}\left(W_k - \mu_k^* T_{\mathrm{slot}}\right)s_{k,n}^* - \zeta^* s_{k,n}^* \\ &\begin{cases} <0 & p_{k,n}^* = 0 \\ =0 & 0 < p_{k,n}^* < p_T \end{cases} \end{aligned} \tag{4.10}$$

and

$$\left.\frac{\partial L}{\partial s_{k,n}}\right|_{p_{k,n}=p^*_{k,n},s_{k,n}=s^*_{k,n}} = \phi_{k,n}(\mu^*_k,\zeta^*) - \theta^*_n$$

$$\begin{cases} < 0 & s^*_{k,n} = 0 \\ = 0 & 0 < s^*_{k,n} < 1 \\ > 0 & s^*_{k,n} = 1 \end{cases} \tag{4.11}$$

where

$$\phi_{k,n}(\mu^*_k,\zeta^*) = \frac{B}{N}\left(W_k - \mu^*_k T_{\text{slot}}\right)\log_2\left(1 + p^*_{k,n}\gamma_{k,n}\right) - \zeta^* p^*_{k,n} \tag{4.12}$$

The optimal PA can be derived from (4.10) and $0 \le p^*_{k,n} \le p_{\text{T}}$ that:

$$p^*_{k,n} = \left[\frac{B(W_k - \mu^*_k T_{\text{slot}})}{N\zeta^*} - \frac{1}{\gamma_{k,n}}\right]^+ \tag{4.13}$$

where $[x]^+ = \begin{cases} x & x > 0 \\ 0 & x \le 0 \end{cases}$.

Substituting (4.13) into (4.12), it can be shown that

$$\phi_{k,n}(\mu^*_k,\zeta^*) = -\left[\frac{B(W_k - \mu^*_k T_{\text{slot}})}{N} - \frac{\zeta^*}{\gamma_{k,n}}\right]^+ \tag{4.14}$$
$$+\frac{B(W_k - \mu^*_k T_{\text{slot}})}{N}\left[\log_2\left(\frac{B\gamma_{k,n}(W_k - \mu^*_k T_{\text{slot}})}{N\zeta^*}\right)\right]^+$$

Considering the original constraints (C4) and (C5), each subcarrier can only be assigned to one user, or saying $s^*_{k,n}$ is either 1 or 0. It can be derived from (4.11) and (4.14) that the nth subcarrier should be assigned to the user by the following rule:

$$s^*_{k,n} = \begin{cases} 1 & k = k^* \\ 0 & k \ne k^* \end{cases} \tag{4.15}$$

where

$$k^* = \underset{k}{\operatorname{argmax}} \left\{ -\left[\frac{B(W_k - \mu_k^* T_{\text{slot}})}{N} - \frac{\zeta^*}{\gamma_{k,n}}\right]^+ + \frac{B(W_k - \mu_k^* T_{\text{slot}})}{N}\left[\log_2\left(\frac{B\gamma_{k,n}(W_k - \mu_k^* T_{\text{slot}})}{N\zeta^*}\right)\right]^+ \right\} \tag{4.16}$$

It is straightforward that $p_{k,n} = 0$ if $s_{k,n} = 0$. Therefore, the optimal PA is:

$$p_{k,n}^* = \begin{cases} \left[\frac{B(W_k - \mu_k^* T_{\text{slot}})}{N\zeta^*} - \frac{1}{\gamma_{k,n}}\right]^+ & k = k^* \\ 0 & k \neq k^* \end{cases} \tag{4.17}$$

The optimal subcarrier and power allocation of (4.4) are concluded in (4.15) and (4.17), respectively. The next step is to find the optimal Lagrange multipliers. Note that (C15) is achieved when (4.15) is satisfied, which means that θ_n can be omitted. Hence, the aim is to determine the optimal Lagrange multipliers μ_k^* and ζ^* satisfying

$$f(\zeta^*) = p_{\text{T}} - \sum_{n=1}^{N}\sum_{k=1}^{K}\left[\frac{B(W_k - \mu_k^* T_{\text{slot}})}{N\zeta^*} - \frac{1}{\gamma_{k,n}}\right]^+ = 0 \tag{4.18}$$

and

$$g(\mu_k^*) = Q_k - \frac{BT_{\text{slot}}}{N}\sum_{n=1}^{N}\left[\log_2\left(\frac{B\gamma_{k,n}(W_k - \mu_k^* T_{\text{slot}})}{N\zeta^*}\right)\right]^+ = 0 \tag{4.19}$$

where μ_k^* and ζ^* are regarded as constants in (4.18) and (4.19), respectively.

The Bisection algorithm [78] is used to search the Lagrange multipliers μ^* and ζ^*. First of all, $f(\zeta)$ and $g(\mu_k)$ are mono-increasing functions. It can be derived that the lower and upper bounds of the feasible search region of μ_k are given by $\mu_k^L = 0$ and $\mu_k^R = \frac{W_k}{T_{\text{slot}}}$, respectively,. Similarly, it can be found that the lower and upper bounds of the feasible search region of ζ are $\zeta_L = 0$ and $\zeta_R = BW_{\max}/p_{\text{T}}$, respectively, where $W_{\max} = \max\{W_k\}$. The search algorithm for the Lagrange

multipliers is presented as follows:

1. Initialize ζ and μ_k;

2. For each user k:
   Update $\mu_k^L = \begin{cases} \mu_k^L & g(\mu_k) > 0 \\ \mu_k & g(\mu_k) < 0 \end{cases}$, $\mu_k^R = \begin{cases} \mu_k^R & g(\mu_k) < 0 \\ \mu_k & g(\mu_k) > 0 \end{cases}$, and $\mu_k = (\mu_k^L + \mu_k^R)/2$; repeat these procedures until $g(\mu_k^*) = 0$

3. If $f(\zeta^*) \neq 0$, adjust $\zeta_L = \begin{cases} \zeta_L & f(\zeta^*) > 0 \\ \zeta & f(\zeta^*) < 0 \end{cases}$ and $\zeta_R = \begin{cases} \zeta_R & f(\zeta^*) < 0 \\ \zeta & f(\zeta^*) > 0 \end{cases}$, and update $\zeta = \frac{\zeta_L + \zeta_R}{2}$;

4. Update μ^* based on Step 2 with ζ obtained in Step 3;

5. Repeat Steps 3 and 4 to obtain μ_k^* and ζ^*, which satisfy both $f(\zeta^*) = 0$ and $g(\mu_k^*) = 0$.

It can be seen from (4.17) that the power is allocated according to the modified water-filling algorithm [39], because the weight is introduced into the objective function (4.4). Since the search for solutions to (4.15) and (4.17) requires a high complexity [8], the optimal RA algorithm is more suitable for the environment where the channels are constant over a large number of time slots. In this thesis, it assumes quasi-static fading channels, which vary in different time slots. Thus, a suboptimal RA algorithm with a lower complexity is desirable.

4.3.2 Suboptimal Resource Allocation

To make a good trade-off between complexity and performance, a suboptimal RA algorithm, which performs SP and PA separately, is proposed.

Firstly the subcarrier allocation is performed, assuming equal power across all subcarriers, and then the PA is conducted. With equal power on each subcarrier, the subcarrier allocation should assign subcarrier n to user k rather than user

j, if $W_k R_{k,n} > W_j R_{j,n}$. After SA, PA is performed by using KKT conditions [79]. Let Φ denote the user index set. The dynamic algorithm to implement the suboptimal MWSC based subcarrier and power allocation is described as follows:

1. Initialization: Set $\Phi = \{1, ..., K\}$, $R_k = 0$, $\Omega_k = \emptyset$ for $\forall k \in \Phi$ and $p_{k,n} = p_T/N$ $(\forall n \in \{1, ..., N\}, \quad \forall k \in \Phi)$.

2. For $n = 1$ to N

   If $\Phi \neq \emptyset$, find $k^*(n) = \underset{k \in \Phi}{\operatorname{argmax}}\{W_k R_k\}$, and assign subcarrier n to user $k^*(n)$; update $\Omega_{k^*(n)} = \Omega_{k^*(n)} \cup n$ and $R_{k^*(n)}$; if $R_{k^*(n)} T_{\text{slot}} \geq Q_{k^*(n)}$, *i.e.*, user $k^*(n)$ has obtained sufficient resources for sending out all data, remove $k^*(n)$ from set Φ.

3. Allocate power to subcarrier n $(\forall n \in \{1, ..., N\})$ by

$$p_{k,n} = \begin{cases} \frac{W_k \left(p_T + \sum_{m=1}^{K} \sum_{q \in \Omega_m} \frac{1}{\gamma_{m,q}} \right)}{\sum_{m=1}^{K} W_m |\Omega_m|} - \frac{1}{\gamma_{k,n}} & k = k^*(n) \\ 0 & k \neq k^*(n) \end{cases} \tag{4.20}$$

4. For $n = 1$ to N

   If $p_{k^*(n),n} \leq 0$, *i.e.*, there is no enough power for subcarrier n which has small $W_{k^*(n),n}$ and/or $\gamma_{k^*(n),n}$, subcarrier n will be "ignored" for PA and removed from set $\Omega_{k^*(n)}$.

5. Repeat Steps 3 and 4 until $p_{k,n} > 0$ $(\forall n \in \{1, ..., N\}, \forall k \in \{1, ..., K\})$.

Note that constraint (C3) is only applied to subcarrier allocation in the proposed suboptimal RA algorithm, which has a dominant impact on the system capacity [3].

Consider users j and k $(j, k \in \{1, ..., K\}, j \neq k)$, and subcarriers m and n $(m \in \Omega_j$ and $n \in \Omega_k)$ which are two arbitrary subcarriers allocated to users j

and k, respectively. Letting $\partial J/\partial p_{k,n} = \partial J/\partial p_{j,m} = 0$, it is easy to have

$$p_{j,m} - p_{k,n} \geq \frac{1}{\gamma_{k,n}} - \frac{1}{\gamma_{j,m}} \tag{4.21}$$

when $W_j \geq W_k$. According to (4.21), it can be derived that $p_{j,m} - p_{k,n} = \frac{1}{\gamma_{k,n}} - \frac{1}{\gamma_{j,m}}$ when $W_j = W_k$. This implies that when two subcarriers have equal weights, the subcarrier with a better channel gain is allocated more power, which is the same as the water-filling [39] result. When $W_j > W_k$, $p_{j,m} - p_{k,n} > \frac{1}{\gamma_{k,n}} - \frac{1}{\gamma_{j,m}}$, which means that the subcarrier corresponding to the user with a higher weight is allocated more power than the case of water-filling.

4.4 Packet Dependent Scheduling

The weights of the MWSC based cross-layer design described in Section 4.2 contain the QoS information, and can be obtained from scheduling at the MAC layer. In this section, two queue based scheduling schemes: M-LWDF [72] and MDU [23], which can be combined with the queue based MWSC cross-layer design using the cost function given by (4.5), are reviewed. Then a packet based scheduling scheme is proposed in Subsection 4.4.3. Notice that all scheduling schemes are performed in every time slot and the weights are updated in every slot as well.

4.4.1 Modified Largest Weighted Delay First Scheduling

The M-LWDF [72] scheduling scheme can keep the delays of most queues below a bound. The work in [72] is extended to the case of OFDM systems. Let U_i and S_i respectively denote the delay tolerance and HOL packet delay for queue i. Define δ_i as the maximum allowed probability that $S_i > U_i$. The weight of queue i is given by:

$$W_i = -\frac{S_i \log(\delta_i)}{U_i \overline{R}_i} \tag{4.22}$$

where $\overline{R}_i$ is the data rate of queue i averaged over slots.

Substituting (4.22) into (4.5), it can be deduced that with M-LWDF, the queue which has a higher HOL packet delay relative to the delay bound (S_i/U_i), a higher instantaneous data rate relative to the average data rate ($R_i/\overline{R}_i$), and a higher requirement for the outage probability (*i.e.*, a larger $-\log(\delta_i)$), is given a higher priority to be served.

4.4.2 Maximum Delay Utility Scheduling

The so-called MDU scheduling scheme in [23] is to maximize the utility functions with respect to the delay. Define $f_i(\overline{S}_i)$ as a decreasing utility function of the average HOL packet delay $\overline{S}_i = \overline{Q}_i/\overline{\lambda}_i$ for queue i [80], where $\overline{S}_i$, $\overline{Q}_i$ and $\overline{\lambda}_i$ denote the short-term average HOL delay, the average queue length and the average arrival rate of queue i calculated in the current slot, respectively. In practice, systems do not serve an empty queue. Thus, it is straightforward to have $\overline{\lambda}_i = \overline{R}_i$ [23] and $\overline{S}_i = \overline{Q}_i/\overline{R}_i$. The average length of queue i is updated during each slot. Letting Q_i denote the length of queue i in current slot, and $\overline{Q}'_i$ denote the average length of queue i calculated in previous slot, the average length of queue i calculated in current slot is given by $\overline{Q}_i = \alpha Q_i + (1-\alpha)\overline{Q}'_i$, where $0 < \alpha < 1$. Similar to M-LWDF [72], MDU is a queue based scheduling scheme, with the weight W_i in (4.5) expressed as:

$$W_i = \frac{-1}{\overline{\lambda}_i} \cdot \frac{\partial f_i(\overline{S}_i)}{\partial \overline{S}_i} \tag{4.23}$$

4.4.3 Packet Dependent Scheduling

The mechanism of the conventional queue based scheduling such as M-LWDF and MDU is to assign the same weight to all packets in a queue, and serve the selected queues until either the data or PHY layer resources are exhausted during each

slot. This however leads to inefficiency if some packets in the unselected queues are more urgent than some packets in the currently served queues. A packet dependent scheduling scheme, which assigns different weights to different packets contained in the same queue of the same user, is now proposed. The packet weights are determined based on the delay, packet size and QoS priority level of the packets. Therefore, it is more flexible and efficient than the queue based scheduling.

Defining $U_{k,i}$ as the delay tolerance for queue i ($i \in \{1, ..., I_k\}$) of user k ($k \in \{1, ..., K\}$), a packet within this queue will be dropped if its waiting time is larger than $U_{k,i}$. In general, the delay tolerance of the voice, variable bit rate (VBR) video and BE traffic queues is set to be low, medium and high, respectively. A guard interval $G_{k,i}$ is introduced to reduce the packet drop rate, which means that the urgent packets belonging to queue i of user k should be given a high priority within the last $G_{k,i}$ duration before they timeout. Letting $t_c \in [0, \infty)$ denote the current time, for the lth packet that belongs to queue i of user k and arrives at time $t_l \in [0, t_c]$, the delay is given by $S_{k,i,l} = t_c - t_l$. Thus, the time left that the packet becomes urgent is denoted by $C_{k,i,l}$ (in msec), which is expressed as:

$$C_{k,i,l} = U_{k,i} - S_{k,i,l} - G_{k,i} \tag{4.24}$$

If $C_{k,i,l} < 0$, *i.e.*, $S_{k,i,l} > U_{k,i} - G_{k,i}$, it means that the packet is urgent, and will timeout within the guard interval of $G_{k,i}$. Hence, the packets with $C_{k,i,l} < 0$ should be given a high priority during this guard interval.

Further, let $D_{k,i,l}$ denote the packet size (in bits) of the packet that belongs to queue i of user k, and arrives at time t_l. Define $\beta_{k,i} \in [1, \infty)$ as the QoS priority level for queue i of user k. In general, the QoS priority levels for the voice traffic queues, video traffic queues and BE traffic queues are high, medium and low, respectively.

$W_{k,i,l}$, the weight of the to-be-served packets that belong to queue i of user k, and arrive at time t_l, is designed by using the following equation:

$$W_{k,i,l} = \begin{cases} \beta_{k,i} D_{k,i,l}/(C_{k,i,l}+1) & (C_{k,i,l} \geq 0) \\ \beta_{k,i} D_{k,i,l} & (C_{k,i,l} < 0) \end{cases} \tag{4.25}$$

According to (4.25), the proposed PD scheduling scheme assigns a higher weight to the packets in a queue which have a higher QoS priority level, a larger data amount and a fewer time left to become urgent. The packets with higher weights will be transmitted first, regardless of whether they are in the same queue or not. Each packet is assigned a unique sequence number, which ensures that it can be reassembled during transmission. Therefore, the proposed PD scheduling scheme is more flexible and efficient than conventional queue based scheduling schemes [23, 72], where all packets in the same queue are assigned the same weight.

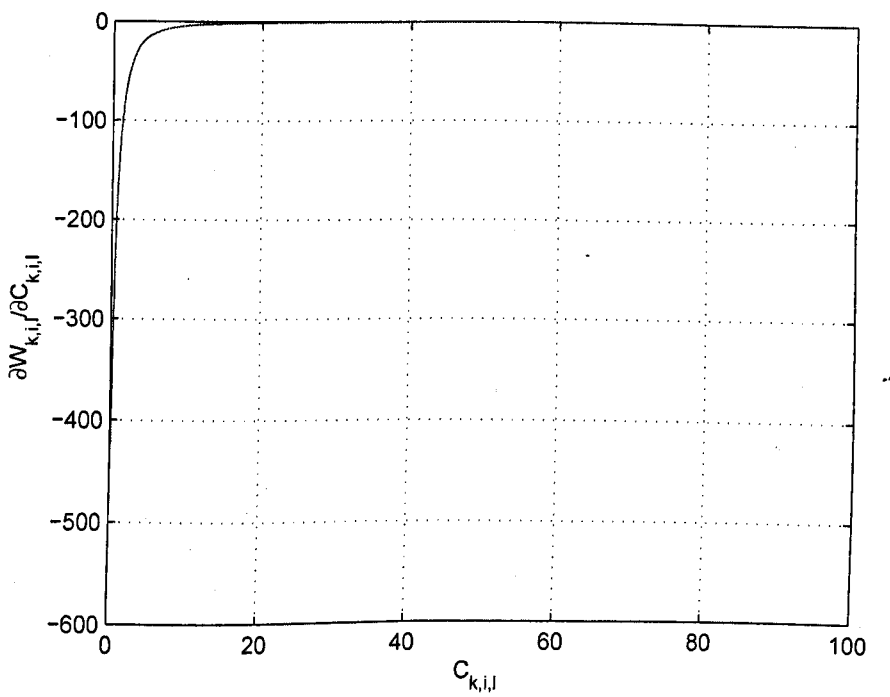

Figure 4.2: The first derivative of $W_{k,i,l}$

1. If $C_{k,i,l} \geq 0$, *i.e.*, $S_{k,i,l} \leq U_{k,i} - G_{k,i}$, the weight $W_{k,i,l}$ in (4.25) increases

with the decrease of $C_{k,i,l}$. In particular, it has:

$$\partial W_{k,i,l}/\partial C_{k,i,l} = -\beta_{k,i} D_{k,i,l}/(C_{k,i,l}+1)^2 \tag{4.26}$$

This implies that even with a small valued $C_{k,i,l}$, *i.e.*, a relatively small packet delay, the variation of the weight $W_{k,i,l}$ is less sensitive to the variation of $C_{k,i,l}$. In this case, the QoS priority level and data amount play a more important role in weight calculation. This can be seen in Figure 4.2, where $\partial W_{k,i,l}/\partial C_{k,i,l}$ approaches zero when $C_{k,i,l} > 20$. On the other hand, the fewer time left for the packet to become urgent, the larger the packet delay and the faster the weight increases.

2. If $C_{k,i,l} < 0$, *i.e.*, $S_{k,i,l} > U_{k,i} - G_{k,i}$, it means that the packets arriving at time t_l will timeout in less than $G_{k,i}$ msec, and should be given a high priority during the guard interval to avoid the drop of packets. In this case, $W_{k,i,l}$ in (4.25) maintains its maximum value, which is determined by the data amount and QoS priority level of the packets. For instance, if two voice packets (of the same QoS priority level) both fall in the guard interval, the packet with a larger data amount will be assigned a higher weight. On the other hand, if two packets are of the same data amount, the packet with a higher QoS priority level will obtain a higher weight.

To make a good trade-off between performance and complexity, only the first $\Lambda_{k,i}$ to-be-served packets (most urgent packets) in queue i of user k are selected for weight calculation of the MWSC based cross-layer design. The $\Lambda_{k,i}$ selected packets are divided into two sets: 1) the packets falling in the guard interval (*i.e.*, to time out within $G_{k,i}$), denoted by $\mathbb{L}^{\mathrm{U}}_{k,i}$; 2) the rest packets, denoted by $\overline{\mathbb{L}^{\mathrm{U}}_{k,i}}$. It can be deduced that $|\mathbb{L}^{\mathrm{U}}_{k,i}| + |\overline{\mathbb{L}^{\mathrm{U}}_{k,i}}| = \Lambda_{k,i}$. Using (4.25), the weight W_k in (4.4)

for the current slot is given by:

$$W_k = \sum_{i=1}^{I_k} \sum_{l \in \mathbb{L}_{k,i}^{U} \cup \overline{\mathbb{L}_{k,i}^{U}}} W_{k,i,l} \tag{4.27}$$
$$= \sum_{i=1}^{I_k} \beta_{k,i} \left[\sum_{l \in \mathbb{L}_{k,i}^{U}} D_{k,i,l} + \sum_{l \in \overline{\mathbb{L}_{k,i}^{U}}} \frac{D_{k,i,l}}{C_{k,i,l}+1} \right]$$

It is important to determine the value of $\Lambda_{k,i}$ in (4.25). With a small-valued $\Lambda_{k,i}$, the complexity decreases, while less QoS information of the to-be-served packets is obtained from the MAC layer for RA. As a result, the performance of the cross-layer design is degraded. In this case, the urgent packets, represented by the first term in (4.25), play a more important role in the weight calculation. On the other hand, with a large-valued $\Lambda_{k,i}$, the complexity increases, while more QoS information of the to-be-served packets is reflected in the weights, which leads to an improved system performance. In practice, a reasonable value of $\Lambda_{k,i}$ is below 100.

The proposed PD scheduling scheme can also be easily combined with the queue based MWSC RA, whose complexity is higher than that of the user based MWSC with heterogeneous traffic, as can be shown in Section 4.6.

4.5 Performance Analysis

System stability plays a critical role in network design [2]. In this section, the stability of the proposed cross-layer PD scheduling scheme is analyzed, assuming infinite buffer size, and no packet drop during transmission, as in [23] and [2].

Given optimization objective(s) and allocation constraints, define $\mathbf{R} = [R_1, ..., R_K]^T$ as an instantaneous transmission rate vector achieved by a RA algorithm. Assuming that the channel state process is ergodic, let $\widetilde{C}$ denote the ergodic (long-term average) capacity region, which consists of the average transmission rate vectors

$E\{\mathbf{R}\}$ achieved by all possible RA algorithms and result in reliable communications [23]. $int(\widetilde{C})$ is defined as the interior of $\widetilde{C}$, which refers to the whole region of $\widetilde{C}$ except for its boundary [2]. Let $\widetilde{S}$ denote the stability region achieved by a scheduling algorithm, which is defined as the set of all possible average arrival rate vectors $E\{\boldsymbol{\lambda}\}$with which the system stability is guaranteed by the scheduling algorithm [2], where $\boldsymbol{\lambda} = [\lambda_1, ..., \lambda_K]^T$ denotes an instantaneous arrival rate vector. The maximum stability region that can be achieved by different scheduling algorithms is identical with $int(\widetilde{C})$ [9]. Therefore, if a system is stable, then $\widetilde{S} \subseteq int(\widetilde{C}) \subset \widetilde{C}$. It means that the average arrival rates locate in the interior of the ergodic capacity region, *i.e.*, $E\{\boldsymbol{\lambda}\} \in int(\widetilde{C})$. In other words, if a system is stable, there exists an average transmission rate vector $E\{\mathbf{R}\} \in \widetilde{C}$ so that $E\{\lambda_k\} < E\{R_k\}$ for all $k \in \{1, ..., K\}$. However, this is a necessary but insufficient condition for system stability.

Let $Q_{k,Z}$ denote the queue length of user k up to the current time slot Z, which is given by

$$Q_{k,Z} = \sum_{i=1}^{I_k} \sum_{l \in \mathbb{L}_{k,i}^{Z}} D_{k,i,l} \tag{4.28}$$

where $\mathbb{L}_{k,i}^{Z}$ denotes the set of packets belonging to queue i, user k, up to current slot Z. It was shown in [81] that a system is regarded as stable if

$$\limsup_{P \to \infty} \frac{1}{P} \sum_{Z=1}^{P} \sum_{k=1}^{K} E\{Q_{k,Z}\} < \infty \tag{4.29}$$

which implies that the upper limit of the average queue length is finite. Substituting (4.29) into (4.28), it can be deduced that a system is stable if

$$\limsup_{P \to \infty} \frac{1}{P} \sum_{Z=1}^{P} \sum_{k=1}^{K} \sum_{i=1}^{I_k} \sum_{l \in \mathbb{L}_{k,i}^{Z}} E\{D_{k,i,l}\} < \infty \tag{4.30}$$

Theorem 4.1: *The proposed cross-layer PD scheduling scheme achieves the*

maximum stability region, i.e., the PD scheduling scheme guarantees the system stability if the average data arrival rates locate in the interior of the ergodic capacity region.

The proof of *Theorem 4.1* is shown in Appendix A. Given the optimization objective in (4.4) and allocation constraints in (C1)-(C5), let $\widetilde{C}$ denote the ergodic capacity region achieved by all possible RA algorithms, including the algorithms presented in Section 4.3. *Theorem 4.1* means that if $E\{\boldsymbol{\lambda}\} \in int(\widetilde{C})$, the proposed PD scheduling guarantees the system stability defined by (4.30). In other words, if there exists an average transmission rate vector $E\{\mathbf{R}\} \in \widetilde{C}$ so that $E\{\lambda_k\} < E\{R_k\}$ for all $k \in \{1, ..., K\}$, the system is stable. Therefore, with PD scheduling, $E\{\boldsymbol{\lambda}\} \in int(\widetilde{C})$, *i.e.*, $E\{\lambda_k\} < E\{R_k\}$ for all $k \in \{1, ..., K\}$, becomes a necessary and sufficient condition for system stability.

In the proof of *Theorem 4.1*, a Lyapunov drift [81] is employed, which is one of the most important methods to deal with the stability issues of queuing networks. Define a Lyapunov function [82] $Y(\mathbf{Q}_Z) = \sum_{k=1}^{K} y(Q_{k,Z})$ as a scalar measure of the queue length in the system, where $\mathbf{Q}_Z = [Q_{1,Z}, ..., Q_{K,Z}]^T$, and $Y(.)$ is a nonnegative function. The Lyapunov drift is given by $E\{Y(\mathbf{Q}_{Z+1}) - Y(\mathbf{Q}_Z)\}$, representing the expectation of the change of the Lyapunov function during two adjacent slots. According to (4.25), the packet weights determined by our proposed PD scheduling scheme contain the queue state information (packet size). The equation (A.8) is shown below again for easy reference:

$$E\left\{Y(\mathbf{Q}_{Z+1}) - Y(\mathbf{Q}_Z)\right\} \leq a - b\sum_{k=1}^{K}\sum_{i=1}^{I_k}\sum_{l\in \mathbb{L}_{k,i}^{Z}} E\left\{W_{k,i,l}\right\} \qquad (a, b \in (0, \infty))$$

It can be derived that the Lyapunov drift becomes negative when the average aggregate weight, which includes the information of the average queue length according to (4.25), increases to a certain bound. The negative Lyapunov drift

Table 4.1: Complexity of cross-layer optimization schemes

	MWSC + M-LWDF [72]	MWSC + MDU [23]	MWSC + PD
Resource allocation	$O(\omega KN)$	$O(\omega KN)$	$O(KN)$
Scheduling	$O(\omega K)$	$O(\omega K)$	$O(\omega\Lambda_{k,i}K)$
Overall	$O(\omega KN)$	$O(\omega KN)$	$O(KN)$ $(\omega \leq N/\Lambda_{k,i})$ $O(\omega\Lambda_{k,i}K)$ $(\omega > N/\Lambda_{k,i})$

means that the queue length decreases to reinstate the system stability. On the other hand, if the weights do not contain any queue information, for example, in the case of equal weight $W_k = 1$ $(k \in \{1, ..., K\})$, the increase of the queue length has no impact on the weights, and therefore has no impact on the Lyapunov drift. Thus, the system stability is not guaranteed by a scheduling approach which does not consider the queue state information.

4.6 Complexity Analysis

In this section, a complexity analysis of the proposed cross-layer MWSC based RA and PD scheduling is provided, in comparison to the existing scheduling schemes reviewed in Section 4.4. Without loss of generality and for simplicity, it assumes that all users have ω queues each, *i.e.*, $I_k = \omega$ $(k \in \{1, ..., K\})$. Thus, there are a total number of ωK queues in the system.

The suboptimal user based MWSC RA algorithm described in Subsection III-B requires K comparisons in each of the N iterations for subcarrier allocation, which requires a complexity of $O(KN)$. PA is performed subsequently, whose complexity is $O(N)$ and can be ignored compared to the complexity of subcarrier allocation. Hence, the complexity for one iteration in suboptimal RA is $O(KN)$. To search for the Lagrange multiplier, the suboptimal RA requires a small number of iterations (2-3 iterations on average), which is independent of the numbers of users and subcarriers. The optimal RA algorithm described in Subsection 4.3.2 requires the same complexity of $O(KN)$ as the suboptimal algorithm for one

iteration, according to (4.15) and (4.17). However, it needs to search for two Lagrange multipliers at the same time, and therefore requires 10 times more iterations on average than the suboptimal RA which requires searching for only one Lagrange multiplier. Due to its significant complexity reduction and little performance loss over the optimal algorithm (shown in Figure 4.5), only the suboptimal RA algorithm is considered in the following of this section.

Compared to the user based MWSC RA, the queue based MWSC RA requires ωK comparisons in each of the N iterations for subcarrier allocation, and therefore its overall complexity is $O(\omega KN)$. This implies that the complexity of the user based MWSC RA algorithm is lower than that of the queue based MWSC RA algorithm with heterogeneous traffic, especially when ω, the number of queues per user, is large.

When the queue based MWSC RA is combined with the M-LWDF scheduling, M-LWDF uses (4.22) once to calculate the weight for each of the ωK queues in each slot, and therefore its complexity is $O(\omega K)$. The complexity of RA is $O(\omega KN)$, as discussed above. By considering only the higher order complexity of RA and scheduling, the overall complexity of MWSC+M-LWDF is $O(\omega KN)$, which grows linearly with the numbers of subcarriers, users, and queues per user. Similarly, the overall complexity of MWSC+MDU is $O(\omega KN)$.

Incorporated with the PD scheduling, the complexity of the user based MWSC RA is $O(KN)$, as discussed above. With the PD scheduling at the MAC layer, according to (4.27), $W_{k,i,l}$ is calculated for at most $\Lambda_{k,i}$ packets in each of the ωK queues. Therefore, the complexity of the PD scheduling is $O(\omega \Lambda_{k,i} K)$. With $\Lambda_{k,i} < N$ in general, it has the following discussion:

1. If $\omega = 1$, *i.e.*, each user has a single traffic queue, the overall complexity of MWSC+PD is $O(KN)$, which is the same as the complexity of MWSC+M-LWDF and MWSC+MDU.

2. If $1 < \omega \leq N/\Lambda_{k,i}$, which is the most likely scenario in practice, RA plays a dominant role in the overall complexity of MWSC+PD, which is $O(KN)$. In this case, the overall complexity of the proposed cross-layer design is roughly independent of the number of queues per user, benefiting from the user based RA. With given numbers of users and subcarriers, MWSC+PD achieves a complexity reduction of approximately ω times over MWSC+M-LWDF and MWSC+MDU, whose overall complexities are $O(\omega KN)$.

3. If $\omega > N/\Lambda_{k,i}$, the overall complexity of MWSC+PD is $O(\omega\Lambda_{k,i}K)$. Therefore, the overall complexity of MWSC+PD is around $N/\Lambda_{k,i}$ times lower than that of MWSC+M-LWDF and MWSC+MDU.

The above analysis is summarized in TABLE 4.1, where the overall complexity of each cross-layer design takes account of only the higher order complexity of RA and scheduling, as it is dominant with the increase of the parameters values [83]. It can be concluded that by properly choosing the number of packets used for weight calculation, MWSC+PD requires a lower overall complexity than MWSC+M-LWDF and MWSC+MDU when there are multiple heterogeneous traffic queues for each user simultaneously. For instance, with $K = 10$ users, $N = 512$ subcarriers, $\omega = 3$ queues per user and $\Lambda_{k,i} = 100$, 75 and 50 for voice, video and BE traffic, MWSC+PD achieves a complexity reduction of around 3 times.

4.7 Simulation Results

Simulation results are used to demonstrate performance of the proposed cross-layer design. It is assumed that the system is with a total transmit power $p_T = 1$ W, a slot duration of $T_{\text{slot}} = 2$ msec, and a total bandwidth of $B = 5$ MHz which is divided into $N = 512$ subcarriers. Three kinds of traffic are assumed:

voice, VBR video and BE traffic queues. The delay tolerance for voice, video and BE traffic is set to be $U_{k,i}^{\text{voice}} = 100$ msec, $U_{k,i}^{\text{video}} = 400$ msec and $U_{k,i}^{\text{BE}} = 1000$ msec ($\forall k \in \{1, ..., K\}, \forall i \in \{1, ..., I_k\}$), respectively. The guard interval $G_{k,i}$ is set to be 10 msec for all types of traffic. To test all the algorithms in the worst case, it is assumed that the packets arrive at an interval of $T_{\text{a}}^{k,i} = 1$ msec ($\forall k \in \{1, ..., K\}, \forall i \in \{1, ..., I_k\}$). The packet arrival rates of voice and BE traffic are constantly $\lambda_{k,i}^{\text{voice}} = 64$ Kbps and $\lambda_{k,i}^{\text{BE}} = 500$ Kbps ($\forall k \in \{1, ..., K\}, \forall i \in \{1, ..., I_k\}$), respectively. The packet arrival rate of the VBR video traffic, $\lambda_{k,i}^{\text{video}}$ ($\forall k \in \{1, ..., K\}, \forall i \in \{1, ..., I_k\}$), follows a truncated exponential distribution in each state, with the maximum, minimum and mean of 420 Kbps, 120 Kbps and 239 Kbps, respectively, where the duration of each state follows an exponential distribution with mean 160 msec. It can be calculated that the average packet sizes for voice, video and BE traffic are 64 bits, 239 bits and 500 bits, respectively. The QoS priority levels for the voice, video and BE traffic are respectively set to be $\beta_{k,i}^{\text{voice}} = 1024$, $\beta_{k,i}^{\text{video}} = 512$ and $\beta_{k,i}^{\text{BE}} = 1$ as a result of testing, to provide the best performance. Based on different requirements of different traffic types on performance and complexity, the number of voice, video and BE packets used for weight calculation in (4.27) are set to be $\Lambda_{k,i}^{\text{voice}} = 100$, $\Lambda_{k,i}^{\text{video}} = 75$ and $\Lambda_{k,i}^{\text{BE}} = 50$ ($\forall k \in \{1, ..., K\}, \forall i \in \{1, ..., I_K\}$), respectively. The channel is modeled to be of six independent Rayleigh fading paths with an exponential delay profile and an RMS delay spread of 0.5 μsec. The SNR is defined as the average ratio of the received signal power to noise power for each user.

4.7.1 Comparison of Resource Allocation Algorithms

Firstly, a simple example is used to demonstrate the advantages of the proposed MWSC based RA over the maximum capacity (MC) [3] and proportional fairness (PF) [38] based RA algorithms. In this comparison, it is assumed that each user has only one queue, and it is for VBR video traffic. An equal data rate condition

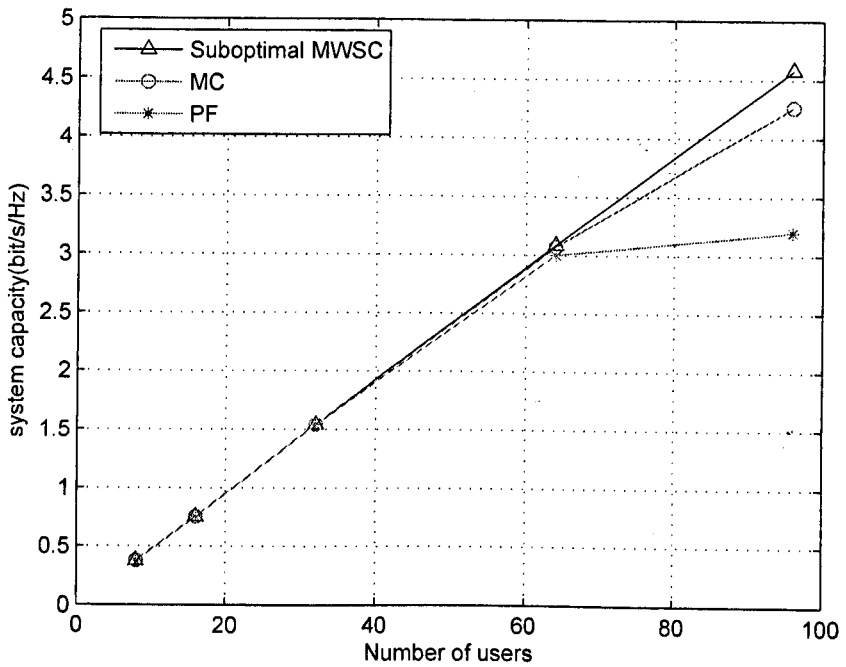

Figure 4.3: System bandwidth efficiency versus the number of users (SNR = 20 dB)

is applied for PF, *i.e.*, $R_1 : ... : R_K = 1 : ... : 1$. To be consistent, all the three algorithms use the PD scheduling scheme at the MAC layer.

As shown in Figure 4.3, all the RA schemes achieve nearly the same capacity with a relatively small number of users. With a large number of users, *e.g.*, 96 users, MWC provides the best performance, with a capacity enhancement of around 7% over MC, benefiting from the condition given by (C3) which avoids any waste in RA. The capacity of PF remains nearly the same with the further increase of the number of users, since the average number of channels per user decreases.

Figure 4.4 shows the average delay of VBR video traffics. MWSC achieves a significantly lower delay than MC and PF, with the increase of the number of users. With 96 users, the delay of MWSC is 60 msec and 270 msec less than that of MC and PF, respectively.

Next, the proposed suboptimal MWSC based RA algorithm is compared with

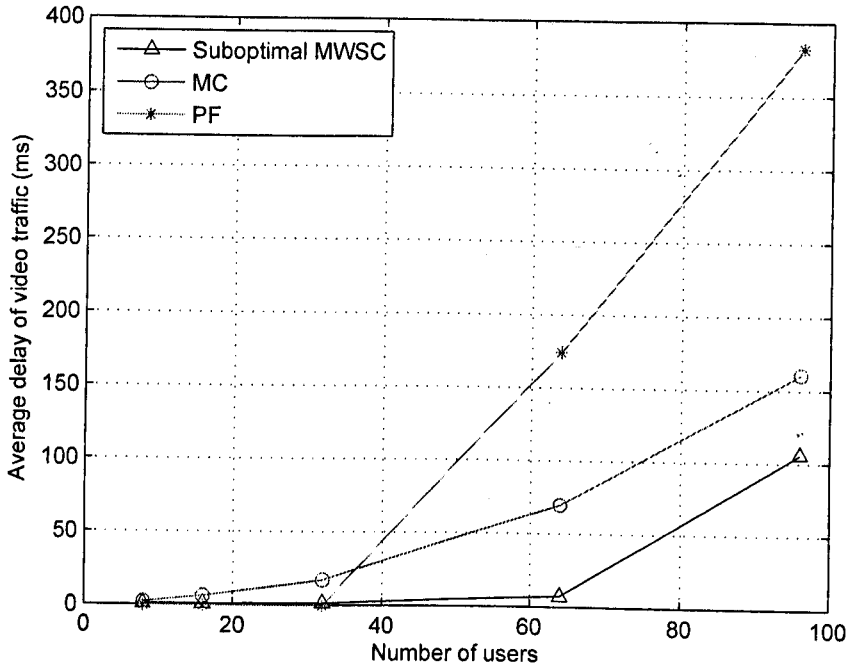

Figure 4.4: Average delays of video traffic with resource allocation algorithms (SNR = 20 dB)

the optimal one. In this case, it is assumed that each user has $I_k = 3$ ($\forall k \in \{1, ..., K\}$) queues: voice, VBR video and BE traffic queues.

Figures 4.5 and 4.6 demonstrate the system throughputs and average delays of the voice and video traffic achieved by using both the optimal and suboptimal RA algorithms proposed in Subsection 4.3.2, and the PD scheduling scheme proposed in Subsection 4.4.3. The optimal RA employs the bisection method [78] to search for the Lagrange multipliers. Compared to the optimal algorithm, the suboptimal RA algorithm provides a similar performance, at a lower complexity, as discussed in Section 4.6.

4.7.2 Comparison of Scheduling Schemes

In the following, a performance comparison of the proposed PD scheduling scheme with M-LWDF and MDU is presented. For MDU, the utility functions defined

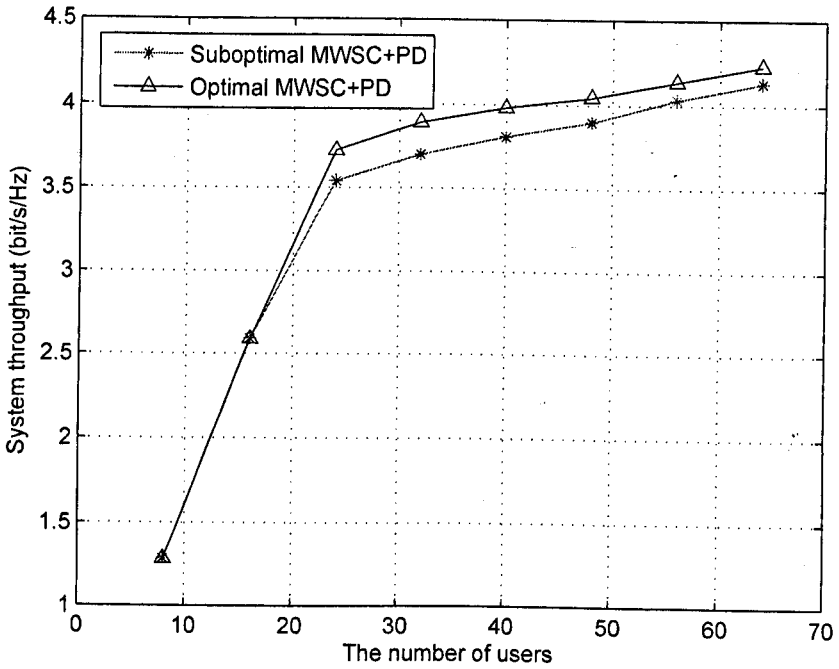

Figure 4.5: Impact of the number of users on the system throughputs of optimal and suboptimal resource allocation schemes (SNR = 20 dB)

in [23] are used. For the purpose of consistency and complexity, the MWSC based suboptimal RA discussed in Subsection 4.3.2 is employed at the PHY layer for all the cross-layer schemes. It is also assumed that each user has $I_k = 3$ ($\forall k \in \{1, ..., K\}$) queues: voice, VBR video and BE traffic queues.

Figures 4.7 to 4.10 demonstrate the impact of the number of users on performance of different cross-layer deigns, with SNR = 20 dB. The system and BE traffic throughputs are shown in Figure 4.7 and Figure 4.8, respectively. PD demonstrates significant performance advantages over MDU and M-LWDF with a wide range of the number of users ($K = 24 \sim 64$ users), while achieving a complexity reduction of about 3 times, as discussed in Section 4.6.

In particular, the system throughput achieved by PD increases with the increase of the number of users, due to enhanced multiuser diversity. With given resources, when the number of users increases, there is a higher degree of freedom for RA, resulting in enhanced multiuser diversity. On the other hand, the increase in

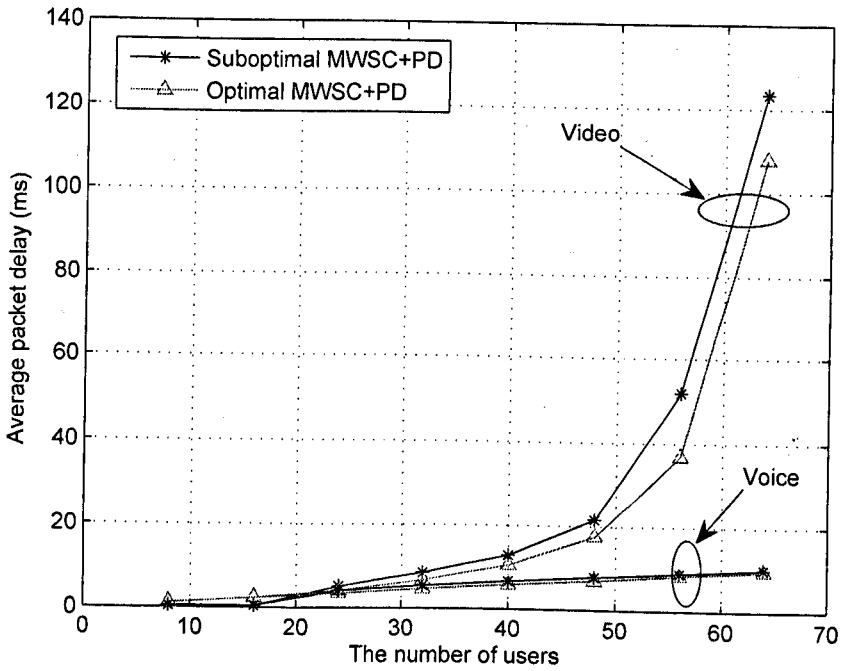

Figure 4.6: Impact of the number of users on average voice and video packet delays of optimal and suboptimal resource allocation schemes (SNR = 20 dB)

the number of users would cause the increase in the weight difference of the HOL packets of various queues, and thus the decrease in the system throughput. This impact on the system throughput is more significant with MDU/M-LWDF than with PD. With PD, the packets with larger weights are served first, whichever queues they belong to. Whereas with MDU/M-LWDF, all packets in a queue which has a larger weight are served first. Therefore, the system throughputs of MDU and M-LWDF benefit less from multiuser diversity with the increase of the number of users in this case. With a large number of users, the system throughput achieved by PD is much higher than that achieved by MDU/M-LWDF. Within the range of small to moderate number of users ($K = 16 \sim 48$), as shown in Figure 4.8, PD outperforms MDU and M-LWDF in terms of the BE throughput. With a large number of users ($K = 56 \sim 64$), the BE throughput achieved by PD is the second highest among the three schemes. Meanwhile, since PD allocates more resources to QoS traffic in this case, it achieves up to two times lower delays

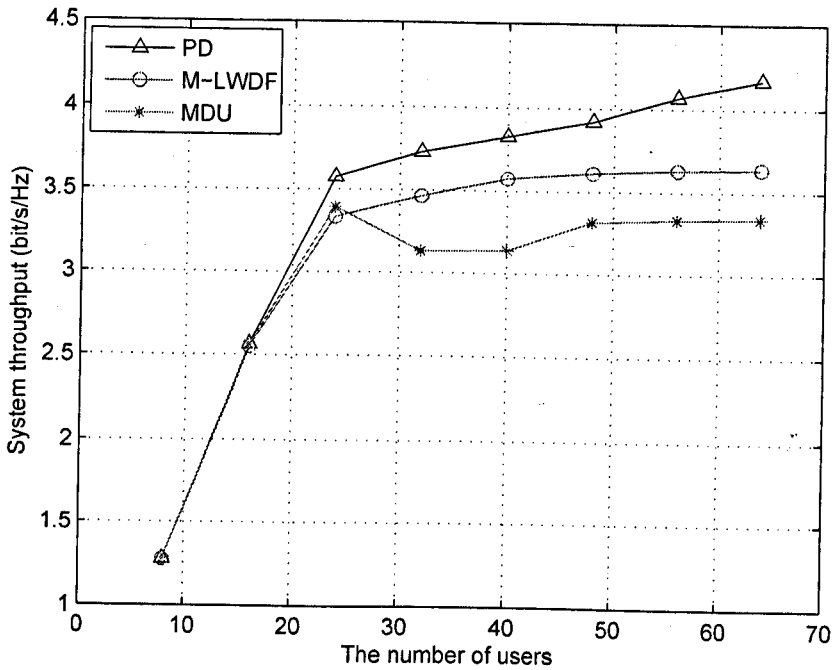

Figure 4.7: Impact of the number of users on system throughput of PD, M-LWDF and MDU scheduling schemes (SNR = 20 dB)

for QoS traffic than M-LWDF and MDU, as can be seen in Figures 4.9 and 4.10.

Figure 4.9 shows the impact of the number of users on the average voice traffic delay of different cross-layer designs. PD achieves a much lower delay than M-LWDF and MDU with a wide range of the number of users ($K = 40 \sim 64$ users). For instance, the average voice delays for both MDU and M-LWDF with $K = 64$ users are about 30 msec, while it is only 10 msec for PD. A similar trend can be observed in Figure 4.10 for the average video traffic delay. With $K = 64$ users, the video delay of PD is around 2 times lower than the delays of MDU and M-LWDF.

The voice and video delay outage probabilities are demonstrated in Figures 4.11 and 4.12, respectively, where $Pr(delay > t)$ denotes the probability that a packet has a delay of greater than t. As can be seen in Figure 4.11, given the delay tolerance of 100 msec, all the voice packets are guaranteed to be transmitted after 30 msec using PD, which is 10 msec less than the case using MDU and M-LWDF.

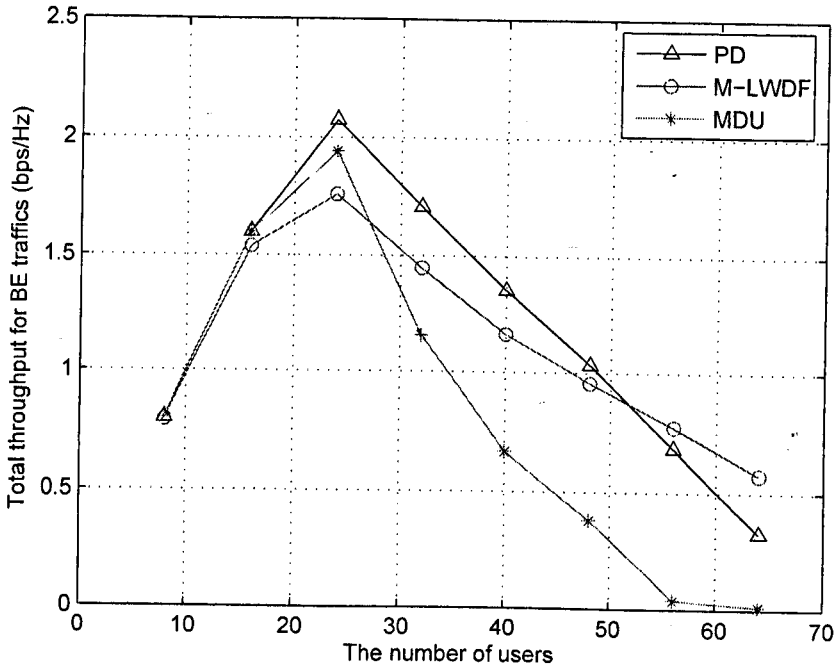

Figure 4.8: Impact of the number of users on BE traffic throughput of PD, M-LWDF and MDU scheduling schemes (SNR = 20 dB)

It is shown in Figure 4.12 that at a delay time of $t = 160$ msec, PD will have had all video packets transmitted, while M-LWDF and MDU still have more than 90% and 80% outstanding packets to be served, respectively. With a delay tolerance of $t = 400$ msec, the video packet drop rates of M-LWDF and MDU are 3.7% and 9%, respectively, while PD results in a zero packet drop rate.

Figures 4.13 to 4.16 show the impact of SNR on performance of different cross-layer schemes with $K = 32$ users. The system throughput is demonstrated in Figure 4.13, where PD outperforms M-LWDF and MDU when the SNR is below 30 dB. At SNR = 15 dB, the bandwidth efficiency achieved by PD is around 12% or 25% higher than those achieved by M-LWDF or MDU, respectively. The throughput of the BE traffic is shown in Figure 4.14. PD outperforms M-LWDF and MDU from a moderate to high SNR range. MDU and M-LWDF achieve a performance similar to that of PD only at a high SNR, when there are sufficient resources for them to transmit all the QoS and BE traffic queues (PD, MDU and

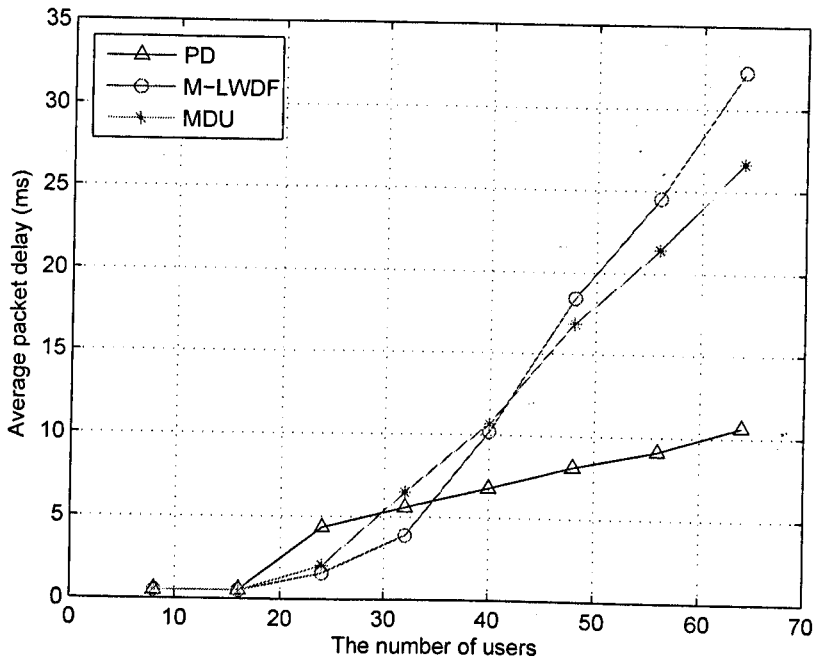

Figure 4.9: Impact of the number of users on average voice packet delays of PD, M-LWDF and MDU scheduling schemes (SNR = 20 dB)

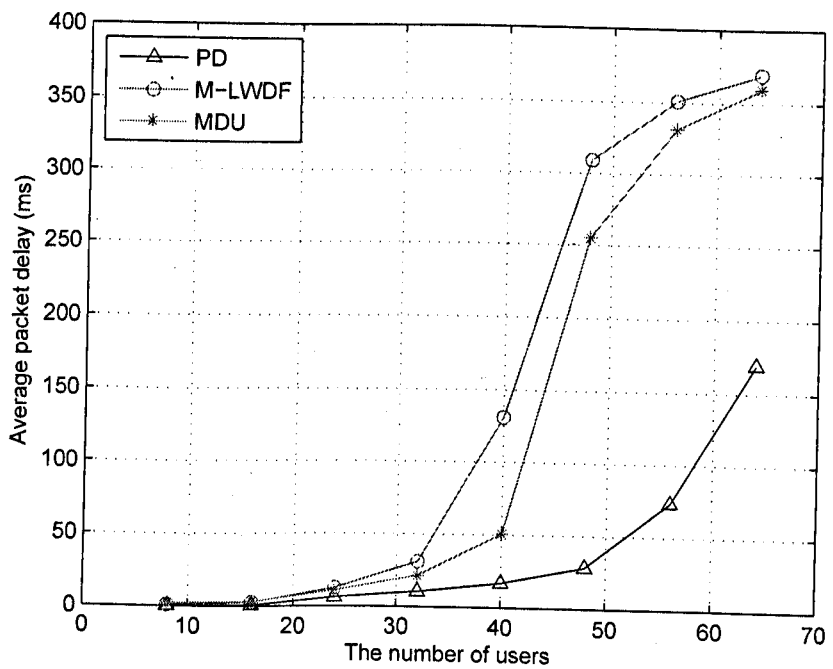

Figure 4.10: Impact of the number of users on average video packet delays of PD, M-LWDF and MDU scheduling schemes (SNR = 20 dB)

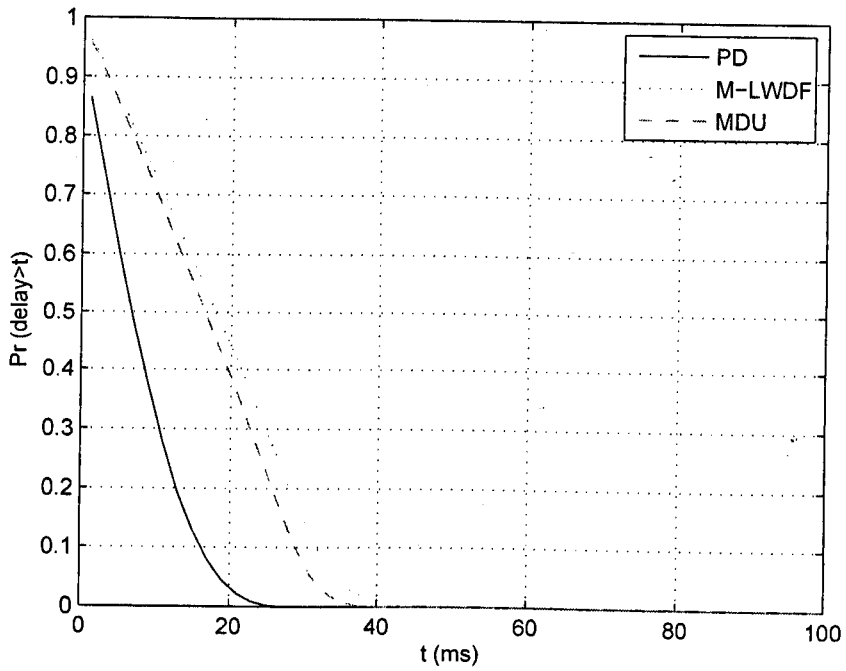

Figure 4.11: Voice traffic delay probability of PD, M-LWDF and MDU scheduling schemes with $K = 48$ users (SNR = 20 dB)

M-LWDF reach the saturation of the BE traffic throughput at SNR = 30 dB).

Figure 4.15 demonstrates the average delay of the voice traffic. Within a moderate to high SNR range, the voice delay of PD is almost the same as those of M-LWDF and MDU. At a low SNR, the voice delay of PD is not as good as those of M-LWDF and MDU. This is because M-LWDF and MDU concentrate limited resources on selected voice queues in each time slot and serve all packets in those queues, resulting in a lower average voice delay than PD, which serves the packets having higher delays in all queues first.

The average delay of the video traffic versus SNR is shown in Figure 4.16, where PD outperforms M-LWDF and MDU when SNR is at $10 \sim 25$ dB. It can also be observed that the higher the SNR, the more advantages of PD over M-LWDF and MDU. At a low SNR, PD achieves a similar video delay compared to M-LWDF and MDU, while providing the lowest packet drop rate, as verified by simulations.

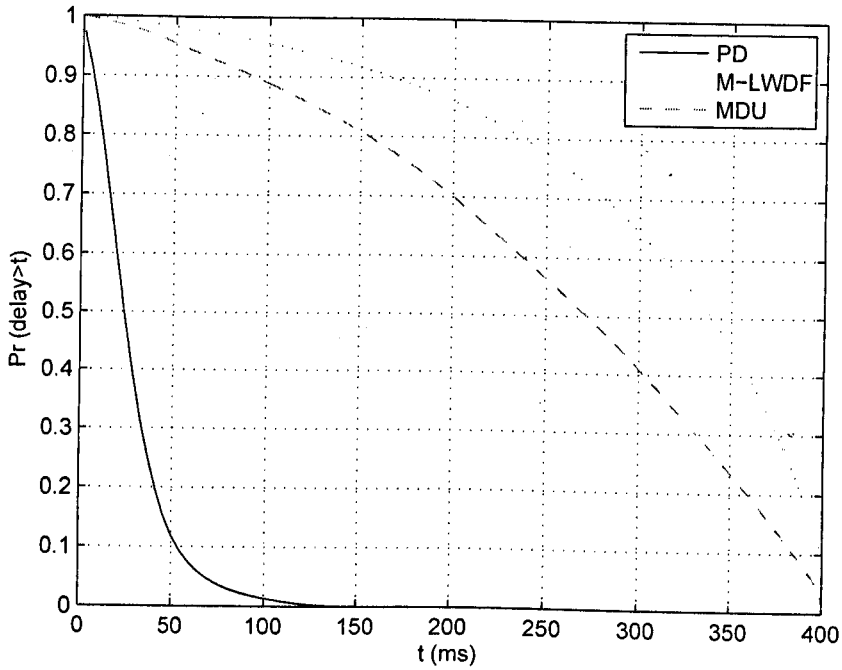

Figure 4.12: Video traffic delay probability of PD, M-LWDF and MDU scheduling schemes with $K = 48$ users (SNR = 20 dB)

It can be deduced from Figures 4.7 to 4.16 that with a wide range of the number of users ($K = 28 \sim 64$ users), and with a moderate to high SNR ($15 \sim 25$ dB), PD outperforms the queue based M-LWDF and MDU in terms of the system throughput, BE traffic throughput, and QoS traffic delays, while achieving a complexity reduction of around 3 times over M-LWDF and MDU, as discussed in Section 4.6.

4.8 Summary

An adaptive cross-layer design with the PD scheduling at the MAC layer has been proposed, for downlink multiuser multitasking OFDM systems with heterogeneous traffic. The proposed suboptimal MWSC based RA algorithm achieves almost the same system performance as the optimal one, at a much lower complexity. Furthermore, it outperforms the MC [3] and PF [38] based RA. The pro-

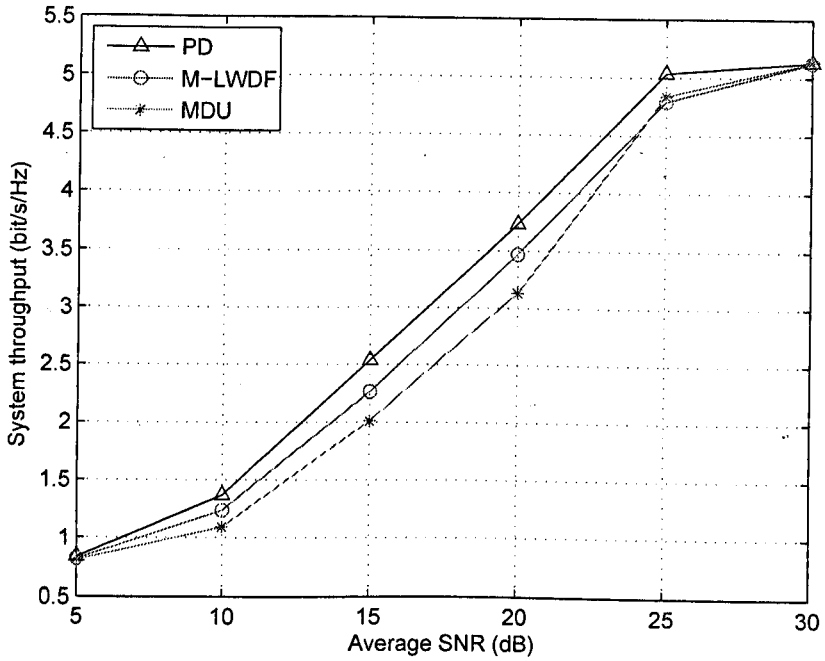

Figure 4.13: Impact of SNR on system throughput of PD, M-LWDF and MDU scheduling schemes with $K = 32$ users

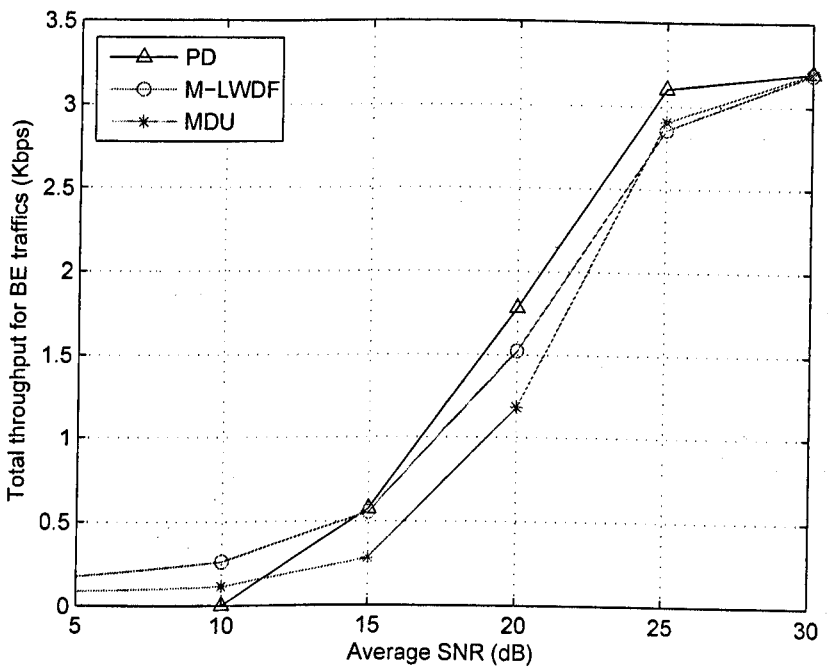

Figure 4.14: Impact of SNR on BE traffic throughput of PD, M-LWDF and MDU scheduling schemes with $K = 32$ users

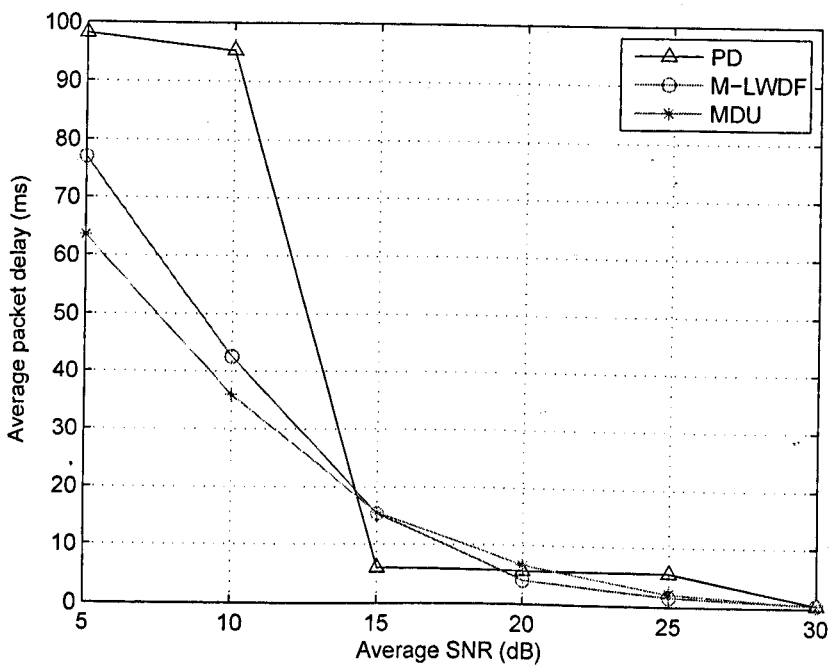

Figure 4.15: Impact of SNR on average voice packet delays of PD, M-LWDF and MDU scheduling schemes with $K = 32$ users

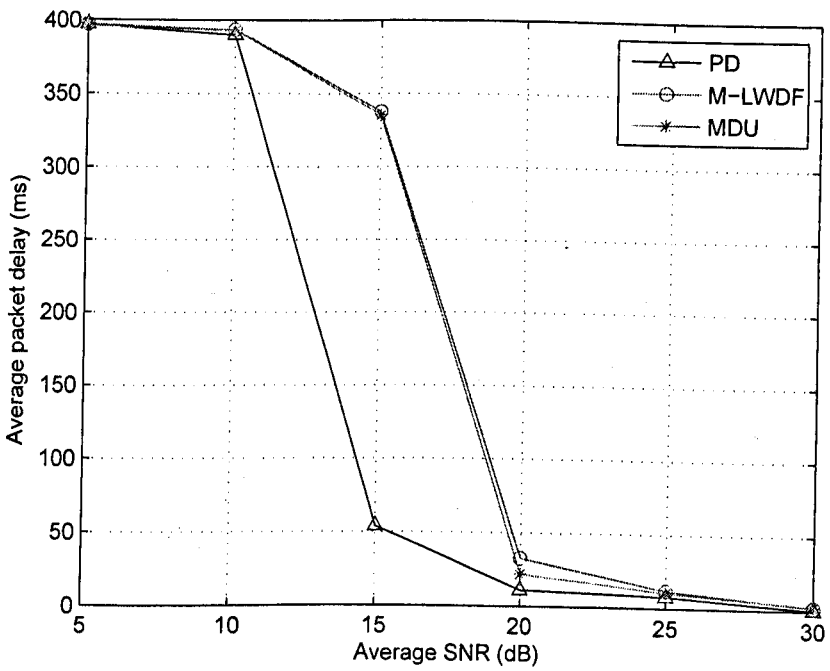

Figure 4.16: Impact of SNR on average video packet delays of PD, M-LWDF and MDU scheduling schemes with $K = 32$ users

posed PD scheduling scheme provides significant performance advantages over the queue based M-LWDF [72] and MDU [23] scheduling schemes in terms of the system throughput, BE traffic throughput, QoS traffic delays and QoS traffic outage probabilities, with a wide range of the number of users, and a moderate to high SNR range. It is also proven to achieve the maximum system stability region. By properly choosing the number of packets used for weight calculation, our user based cross-layer design requires a much lower overall complexity than conventional queue based designs.

Chapter 5

Resource Management for Relay based LTE-Advanced/WiMAX 2 Systems

In 4G wireless communication systems, such as LTE-Advanced/WiMAX 2 systems, base stations are designed to obtain a larger coverage than currently deployed wireless communication systems. Therefore, relays are utilized to help communications between base stations and mobile users.

In the OFDM based relay system, the straight communications between the source and the destinations are replaced with two-phase communications between the source and the relay and between the relay and the destination(s). RA has a significant impact on the system performances, such as the system capacity [43, 49], the outage probability [84] and the BER [85].

In this chapter, an optimal adaptive RA scheme for the multi-destination regenerative relay based LTE-Advanced/WiMAX 2 systems is proposed. The work is different from other existing works [41][42][43][44] in that a practical downlink OFDM based relay network is considered. The relay is used to help the source (base station) to communicate with multiple out-of-range destinations

(users) rather than a single destination (user). The optimal RA scheme consists of three parts: SA, SP and joint source-relay PA. It is shown that the proposed SA at the relay, which was hardly considered in the previous work, and joint source-relay PA plays a dominant role in the three steps of RA. It is also demonstrated that the system capacity benefits more from the multiuser diversity with a shorter distance between the source and the relay. By employing SA and joint source-relay PA only, a near-optimal performance can be achieved. This work can be extended to the multi-hop, multi-relay, and multi-destination uplink/downlink OFDM system.

Following that, an optimal asymmetric RA (ARA) scheme for the multi-destination regenerative relay based LTE-Advanced/WiMAX 2 systems is proposed. ARA is different from the existing symmetric RA schemes in the following aspects. First, it assumes that the bits transmitted over the same subcarrier from the source to the relay may be distributed over different subcarriers when being forwarded from the relay to the destination, unlike the previous work. This increases the degree of freedom for transmission, and more importantly, the amount of feedback information is independent of the number of subcarriers, leading to a dramatic overhead reduction over existing symmetric schemes. Second, the transmission durations at the source and the relay are designed to be asymmetric, which also enhances the degree of freedom for transmission, compared to the previous work assuming equal transmission durations for two consecutive time slots. Simulation results show that the proposed ARA scheme achieves a higher system capacity than existing symmetric RA schemes in [86, 43, 46, 49], with a fast convergence speed. The impact of channel estimation errors on the system capacity, and the impacts of the number of mobile users and distances on time allocation are also investigated.

Section 5.1 first describes the system model for the relay systems investigated in the following sections. The proposed two RA algorithms for symmetric and

asymmetric relay systems are presented in Section 5.2 and 5.3, respectively. Section 5.4 provides the summary.

5.1 System Model

A TDD downlink OFDM relay system is considered, where a source (base station) communicates with destinations (users) via relays. To reduce the cost, operators usually set relay stations apart from each other to reduce the overlaps of the coverage of relays, and each destination can hear from only one relay. As a result, a single-source, multi-relay and multi-destination system can be simply regarded as a combination of a number of subsystems, where each subsystem consists of one source, one relay and a number of destinations under the coverage of the relay. Therefore, without loss of generality, a relay subsystem is considered in this chapter, where a source communicates with K destinations via a relay.

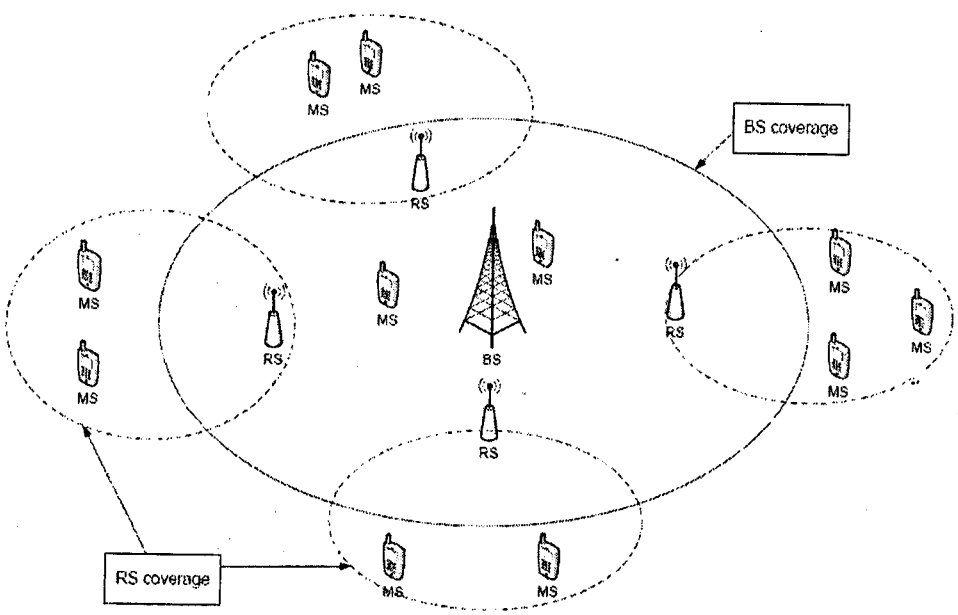

Figure 5.1: Wireless relay network system model (BS: base station; RS: relay station; MS: mobile station)

A total bandwidth of B, which is divided into N independent subcarriers, is allocated to the subsystem. It is assumed that each subcarrier is allocated to

only one relay or destination. It is also assumed that destinations are out of reach of the source for distance or obstacles, which means only the relay can receive and retransmit the signal from source to destinations, and the relay operates in a regenerative strategy. For practical reasons, the system is half-duplex, where each node can not transmit and receive at the same time on the same subcarrier [87]. Hence, communications between the source and destinations cover two time slots. In the first slot, the source sends the data to the relay. The relay receives and decodes the signal. In the second slot, the relay re-encodes the signal and forwards it to the destination. At last, the destination decodes the signal from the relay.

Throughout this chapter, the PA for the mth subcarrier in the first slot is denoted by p_{SR}^m, and the power on subcarrier n allocated to destination k in the second slot is $p_{RD_k}^n$. Let the single-sided power spectral density of AWGN N_0 be equal at relay and destinations, h_{SR}^m and $h_{RD_k}^n$ denote the channel gains for subcarriers m, and subcarrier n allocated to destination k in the two consecutive slots. It can be derived that the CNRs for the first and second slots are $\gamma_{SR}^m = |h_{SR}^m|^2/(N_0B/N)$ and $\gamma_{RD_k}^n = |h_{RD_k}^n|^2/(N_0B/N)$, respectively. It is worth noting that, in our system, pilots only need to be sent out by the relay. Both the base station and the destination can receive the pilots, and use them to estimate the relay-to-source channel h_{RS}^m and the relay-to-destination channel $h_{\mathrm{RD}_k}^n$, respectively. $\hat{h}_{\mathrm{SR}}^m$ can be obtained from the estimate of the relay-to-source channel h_{RS}^m, using reciprocity. The estimates $\hat{h}_{\mathrm{SR}}^m$ and $\hat{h}_{\mathrm{RD}_k}^n$ are then sent back to the relay.

The instantaneous capacities of the two consecutive slots are given by

$$C_{SR}^m = \frac{1}{2}\log_2\left(1 + p_{SR}^m \gamma_{SR}^m\right) \tag{5.1}$$

and

$$C_{RD_k}^n = \frac{1}{2}\log_2\left(1 + p_{RD_k}^n \gamma_{RD_k}^n\right) \tag{5.2}$$

5.2 Symmetric Resource Allocation

The RA algorithms have been investigated for the three-node (single source, single relay and single destination) system [41, 42, 86, 43], which is however not practical. In this section, an optimal adaptive RA scheme for the OFDM based multi-destination regenerative relay system is proposed. The work is different in that a practical downlink OFDM based relay network is considered, where the relay is used to help the source (base station) to communicate with multiple out-of-range destinations (users), rather than a single destination (user). The optimal RA scheme consists of three parts: SA, SP and joint source-relay PA. It is shown that the proposed SA at the relay, which was hardly considered in the previous work, and joint source-relay PA plays a dominant role in the three steps of RA. It is also demonstrated that the system capacity benefits more from the multiuser diversity with a shorter distance between the source and the relay. By employing SA and joint source-relay PA only, a near-optimal performance can be achieved. Our work can be extended to the multi-hop, multi-relay, and multi-destination uplink/downlink OFDM system.

5.2.1 Problem Formulation

It is assumed that $(m,n)_k$ is a subcarrier pair dedicated to communication between the source and destination k ($\forall k \in \{1,...,K\}$). The bits received by the relay on subcarrier m are forwarded to the kth destination over subcarrier n [43]. The aggregate power for subcarrier pair $(m,n)_k$ is given as $p_k^{m,n} = p_{SR}^m + p_{RD_k}^n$. The capacity through the subcarrier pair $(m,n)_k$ to destination k during the two

slots is expressed as:

$$C_k^{m,n} = \min\left\{C_{SR}^m, C_{RD_k}^n\right\} \tag{5.3}$$

where $\min\{x, y\}$ takes the smaller value of x and y.

It is to perform the optimal RA resulting in the highest system capacity. Therefore, the problem is formulated as

$$\max_{\Omega_k, p_{SR}^m, p_{RD_k}^n} \left(\sum_{k=1}^{K} \sum_{(m,n)_k \in \Omega_k} C_k^{m,n} \right) \tag{5.4}$$

subject to (C1) $\sum_{m=1}^{N} p_{SR}^m + \sum_{k=1}^{K} \sum_{n=1}^{N} p_{RD_k}^n \leq p_T$, (C2) $p_{SR}^m > 0$ and (C3) $p_{RD_k}^n > 0$, where Ω_k is the set of subcarrier pairs allocated to destination k and p_T is the total system power.

Lemma 5.1: *Objective function (5.4) can be maximized only when C_{SR}^m and $C_{RD_k}^n$ are equal.*

Proof: It is assumed that there exists a subcarrier pair $C_i^{g,h}$ satisfying $C_{SR}^g \neq C_{RD_i}^h$ in (5.4), while $C_{SR}^m = C_{RD_k}^n$ holds for other subcarrier pairs $(m,n)_k$ ($\forall k \in \{1, ..., K\}$ and $k \neq i$). Without loss of generality, it assumes that $C_{SR}^g > C_{RD_i}^h$ for subcarrier pair $C_i^{g,h}$.

Some power from p_{SR}^g can be shifted to $p_{RD_i}^h$ so that the updated capacities are $\overline{C}_{SR}^g = C_{SR}^g - \Delta_1$, $\overline{C}_{RD_i}^h = C_{RD_i}^h + \Delta_2$, and $\overline{C}_{SR}^g = \overline{C}_{RD_i}^h$ ($\Delta_1 > 0$ and $\Delta_2 > 0$), while keeping the powers of all other subcarrier pairs unchanged. Hence, it has $\overline{C}_i^{g,h} = C_i^{g,h} + \Delta$ ($\Delta = \min\{\Delta_1, \Delta_2\}$), and capacities of other subcarrier pairs remain the same, leading to a higher overall system capacity in (5.4). In other words, the system capacity can always be increased if there is one subcarrier pair with $C_{SR}^m \neq C_{RD_k}^n$. Therefore, (5.4) is maximized only when all the subcarrier pairs satisfy $C_{SR}^m = C_{RD_k}^n$. □

By using *Lemma 5.1*, it can be shown that when p_{SR}^m and $p_{RD_k}^n$ satisfy:

$$p_{SR}^m = \frac{\gamma_{RD_k}^n}{\gamma_{SR}^m + \gamma_{RD_k}^n} p_k^{m,n} \tag{5.5}$$

$$p_{RD_k}^n = \frac{\gamma_{SR}^m}{\gamma_{SR}^m + \gamma_{RD_k}^n} p_k^{m,n} \tag{5.6}$$

the maximum value of (5.3) is obtained as

$$C_k^{m,n} = \frac{1}{2} \log_2 \left(1 + a_k^{m,n} p_k^{m,n}\right) \tag{5.7}$$

where $a_k^{m,n} = \frac{\gamma_{SR}^m \gamma_{RD_k}^n}{\gamma_{SR}^m + \gamma_{RD_k}^n} = \frac{N|h_{SR}^m|^2 |h_{RD_k}^n|^2}{N_0 B(|h_{SR}^m|^2 d_{RD_k}^\alpha + |h_{RD_k}^n|^2 d_{SR}^\alpha)}$. Substituting (5.7) into (5.4), the optimization problem is reformulated as

$$\max_{\Omega_k, p_{SR}^m, p_{RD_k}^n} \sum_{k=1}^{K} \sum_{(m,n)_k \in \Omega_k} \frac{1}{2} \log_2 \left(1 + a_k^{m,n} p_k^{m,n}\right) \tag{5.8}$$

subject to (C4) $\sum p_k^{m,n} \leq p_T$ and (C5) $p_k^{m,n} > 0$.

5.2.2 Optimal Symmetric Resource Allocation

The optimization problem in (5.8) is a combinatorial and convex optimization problem, which is NP-hard [8]. In this subsection, algorithms for SA, SP and joint source-relay PA, and the optimal RA scheme for a multi-destination OFDM based relay system are proposed.

5.2.2.1 Subcarrier Allocation

In this system, the CNR values in the first slot are independent of destinations, while the CNR values in the second slot vary for different destinations.

Proposition 1: If each subcarrier in the second slot is allocated to the destination with the best channel gain on that subcarrier, *i.e.*, subcarrier n will be allocated to user k^*, where $k^* = \underset{k}{\text{argmax}} \gamma_{RD_k}^n$, the maximum system capacity is

achieved.

The multiuser diversity is introduced into the system by *Proposition 1.* As a result, the system capacity increases with the increase in the number of destinations K.

5.2.2.2 Subcarrier Pairing

After SA, each of the N subcarriers in the second slot is associated with a specific destination. To allow the bits transmitted via a subcarrier in the first slot to be forwarded to a subcarrier in the second slot, the following SP algorithm is proposed.

Proposition 2: If the subcarrier of the mth (for all $m \in \{1, ..., N\}$) highest CNR value in the first slot is paired with its counterpart, the subcarrier of the mth highest CNR value in the second slot, the maximum capacity is achieved.

5.2.2.3 Joint Source-Relay Power Allocation

After SA and SP, each subcarrier pair $(m, n)_k$ ($\forall m, n \in \{1, ..., N\}$ and $\forall k \in \{1, ..., K\}$) is determined. Obviously, (5.8) becomes a convex optimization problem [79], and can be solved by applying the Lagrange Multiplier technique.

Proposition 3: After SA and SP, the aggregate power allocated to each subcarrier pair to maximize the system capacity is obtained by using the water-filling algorithm [39], *i.e.*, $p_k^{m,n} = [\frac{B}{N\lambda} - \frac{1}{a_k^{m,n}}]^+$, where λ is the Lagrange multiplier.

After obtaining the aggregate power for each subcarrier pair by *Proposition 3*, the power allocated to each subcarrier at the source and the destination can be obtained by using (5.5) and (5.6).

5.2.2.4 Optimal Resource Allocation

Theorem 5.1: *The optimal RA scheme for a multi-destination OFDM based relay system is in three steps: 1) SA, where the subcarriers for the second slot are*

allocated to multiple destinations; 2) SP, where the subcarriers in two consecutive slots are paired according to their channel gains to serve multiple destinations; 3) joint source-relay PA, where the power is allocated to subcarriers in two slots, subject to the fixed total transmit power.

Note that the proposed solution in *Theorem 5.1* is optimal though the three steps SA, SP and joint source-relay PA are performed in serial. This is because that the three steps are independent of each other. No other RA results, whether they are obtained via joint or in-serial RA, will achieve a higher system capacity than the proposed 3-step RA scheme does. The detailed proof is shown in Appendix B.

The proposed RA scheme can be extended to a multi-hop multi-destination system as follows. Assuming a multi-hop and multi-destination relay system with L hops, the estimated CNR and the allocated power of the n_lth ($\forall n_l \in \{1, ..., N\}$) subcarrier at the lth ($\forall l \in \{1, ..., L\}$) hop are denoted by $\gamma_l^{n_l}$ and $p_l^{n_l}$, respectively. The optimal RA scheme for a multi-hop multi-destination system consists of three steps: 1) SA, where the subcarriers at the last hop are allocated to multiple destinations as in *Proposition 1*; 2) subcarrier grouping, where the N subcarriers for each hop are ordered, and the subcarriers of the same order over L hops are organized into one group to serve the same destination; 3) joint source-relays PA, where the total transmit power for each subcarrier group is determined by the water-filling algorithm [39], subjected to fixed total transmit power. The transmit power allocated to each subcarrier within the group is proportional to the subcarrier carrier power.

5.2.3 Simulation Results

Simulation results are used to show the performance of the proposed RA scheme. It is assumed that a total transmit power of $p_T = 1$ W from the source (base station) and the relay, and all destinations are at the same distance to the relay.

The total bandwidth is $B = 5$ MHz, which is divided into $N = 512$ subcarriers. The channel has six independent Rayleigh fading paths with an exponentially delay profile and an RMS delay spread of 0.5 μs. A fixed total distance of 1000 m between the source and the destinations is assumed, *i.e.*, $d_{SR} + d_{RD_k} = 1000$ m ($\forall k \in \{1, ..., K\}$). The PL exponent is set to be $\tau = 3.5$ as in [88], and the PL model is given by $PL(d) = 38.4 + 35 \log_{10}(d)$ dB [88], as a function of distance d in meter (50 m $< d <$ 5000 m), and the noise power spectral density is $N_0 = -174$ dBm/Hz.

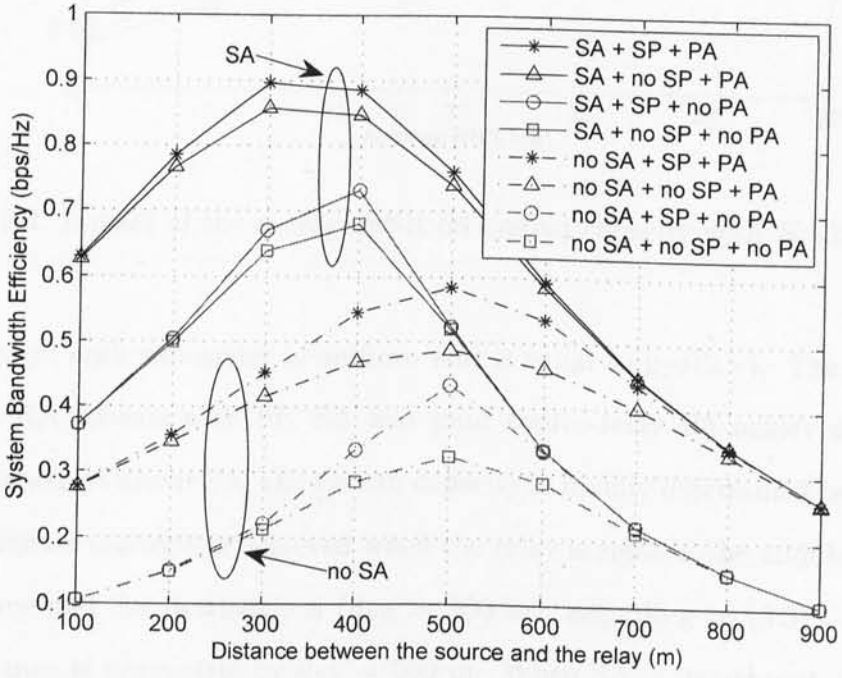

Figure 5.2: Impact of the distance between source and relay on system capacity with $K = 32$ destinations, and a fixed distance of 1000 m between source and destinations

Figure 5.2 demonstrates the impact of the distance between the source and the relay on the system capacity, with $K = 32$ destinations. In this figure, 'no SP' denotes that subcarriers in two consecutive slots are simply paired by their corresponding subcarrier indices. 'no SA' indicates that the subcarriers in the second slot are allocated to destinations randomly. 'no PA' means that the power

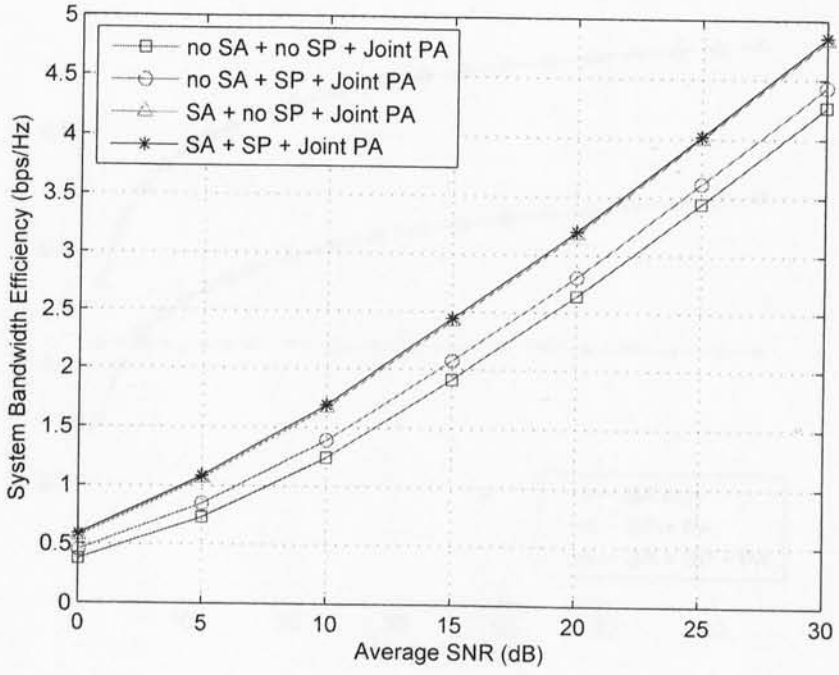

Figure 5.3: Impact of the average SNR on system capacity with K=32 destinations

allocated to each subcarrier is uniform and is equal to $p_T/(2N)$. The proposed optimal RA scheme with SP, SA and joint source-relay PA achieves the best performance. Without SA, the system capacity is mainly determined by PL, and the maximum capacity is achieved when the relay is right in the middle between the source and the destinations (d_{SR} = 500 m), according to (5.3). Also, the performance is symmetric by d_{SR} = 500 m. When SA is introduced, multiuser diversity gains can be achieved in the second slot. Thus, the maximum capacity, as the result of a good balance between PL and multiuser diversity, is achieved at d_{SR} = 300 m. It is interesting to notice that the capacity achieved by SA + SP + PA at d_{SR} = 100 m is around 0.35 bps/Hz higher than that achieved at d_{RD_k} = 100 m ($\forall k \in \{1, ..., K\}$). This is because the multiuser diversity gain increases with the increase of the distance between relay and destinations, *i.e.*, the decrease of d_{SR}. The multiuser diversity introduced by SA leads to a much higher capacity than the case without SA, when the relay is closer to source than

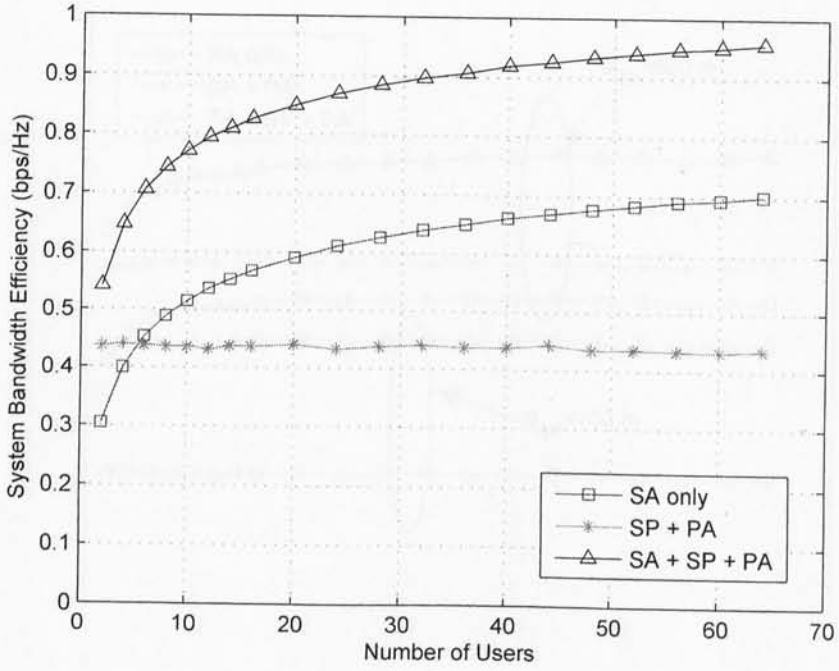

Figure 5.4: Impact of the number of destinations on system capacity with $d_{SR} =$ 300 m, and a fixed distance of 1000 m between source and destinations

to destinations, *i.e.*, $d_{SR} < 500$ m. As can be seen in Figure 5.2, the impact of joint source-relay PA on capacity $C_k^{m,n}$ is also significant. For example, the system capacities achieved by the RA schemes with joint source-relay PA are up to 50% higher than the cases with uniform PA. Therefore, both SA and joint source-relay PA play an important role in the system capacity. By employing SA and joint source-relay PA only, a performance close to the optimal case is achieved, with less complexity. SP only demonstrates an impact on the performance when SA is not employed.

The system capacity versus the average SNR is depicted in Figure 5.3, where both the average SNRs for the two consecutive slots are set to be $\text{SNR} = \text{SNR}_S = \text{SNR}_R$, which means the relay remains an equal distance to the source and the destinations. The number of destinaionts is $K = 32$. As can be seen, the system capacity increases almost linearly with the increase of SNR when $SNR > 5$ dB. Similar to Figure 5.2, the suboptimal RA scheme, consisting of SA and joint

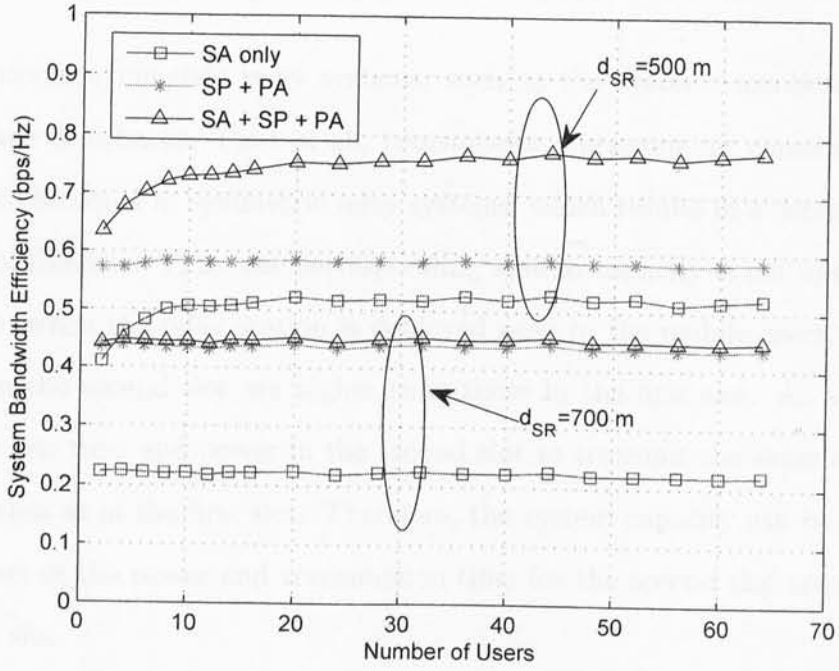

Figure 5.5: Impact of the number of destinations on system capacity with a fixed distance of 1000 m between source and destinations

source-relay PA only, provides a near-optimal performance.

The impact of the number of destinations on the system capacity is depicted in Figures 5.4 and 5.5, with different distances between the source and the relay. The curves were placed in separate figures simply for better visibility. It can be seen that the proposed RA scheme with SA, SP and joint source-relay PA outperforms the SA only scheme in [45] and the SP and PA only scheme in [43] with different values of d_{SR}. With $d_{SR} = 300$ m, the capacity of SA + SP + PA increases rapidly with the increase of the number of destinations. While with $d_{SR} \geq 500$ m, the performance enhancement is insignificant and only occurs when $K < 12$. The SA only scheme provides a better performance than the SP and PA only scheme when $d_{SR} < 500$ m, demonstrating the dominant impact of multiuser diversity when the relay is closer to the source than to the destinations.

5.3 Asymmetric Resource Allocation

Conventional symmetric relay systems, such as the systems mentioned above, have some drawbacks. First of all, transmission durations for consecutive slots are set to be equal in symmetric relay systems, which results in a decrease in the degree of freedom. Thus, the corresponding system capacity is not optimal. For instance, when the relay station is deployed near to the mobile users, the CNR values in the second slot are higher than those in the first slot. As a result, it requires less time and power in the second slot to transmit the same amount of information as in the first slot. Therefore, the system capacity can be increased when part of the power and transmission time for the second slot are shifted to the first slot.

Secondly, the symmetric RA schemes in [43, 86] assumed that the bits transmitted over a subcarrier from the source to the relay must be forwarded from the relay to the destination as a group over a specific subcarrier, which further reduces the degree of freedom for transmission.

Furthermore, the symmetric RA schemes in [43] requires that the CSI for all N subcarriers in the second slot be sent from the relay to the source. As a result, the amount of the feedback information increases with the number of subcarriers. With a large number of subcarriers, this method requires a tremendous spectrum overhead.

In order to overcome the drawbacks of the symmetric RA schemes, the optimal asymmetric RA scheme, referred to as ARA, is now proposed. ARA is different from the previous work in the following aspects. First, the bits transmitted over the same subcarrier from the source to the relay may be distributed over different subcarriers when being forwarded from the relay to the destination with the proposed ARA, unlike the previous work. Secondly, the transmission durations at the base station and the relay station are not necessarily equal, which enhances

the degree of freedom for transmission.

5.3.1 Problem Formulation

In asymmetric relay systems, the instantaneous capacities of the two consecutive slots C_{SR}^m and $C_{RD_k}^n$ in (5.1) and (5.2) become

$$C_{SR}^m = \frac{T_1}{T}\log_2\left(1 + p_{SR}^m \gamma_{SR}^m\right) \tag{5.9}$$

and

$$C_{RD_k}^n = \frac{T_2 \beta_{RD_k}^n}{T}\log_2\left(1 + p_{RD_k}^n \gamma_{RD_k}^n\right) \tag{5.10}$$

where $\beta_{\mathrm{RD}_k}^n$ is a binary value to indicate whether the nth relay-to-destination subchannel is occupied by the kth mobile user. When subchannel n is employed in the second slot from the relay to mobile user k, it has $\beta_{\mathrm{RD}_k}^n = 1$, otherwise $\beta_{\mathrm{RD}_k}^n = 0$.

Given fixed total transmission power p_T and fixed total transmission duration T, the aim is to maximize the system capacity from the base station to multiple mobile users, which is the minimum of the capacities achieved in the two consecutive slots. The problem is formulated as

$$\underset{p_{\mathrm{SR}}^m, p_{\mathrm{RD}_k}^n, T_1, T_2, \beta_{\mathrm{RD}_k}^n}{\operatorname{argmax}} \left(\min\left\{\sum_{m=1}^{N} C_{\mathrm{SR}}^m, \sum_{k=1}^{K}\sum_{n=1}^{N} C_{\mathrm{RD}_k}^n\right\}\right) \tag{5.11}$$

subject to

$$\sum_{m=1}^{N} p_{\mathrm{SR}}^m + \sum_{k=1}^{K}\sum_{n=1}^{N} p_{\mathrm{RD}_k}^n \le p_T, \quad p_{\mathrm{SR}}^m > 0, \quad p_{\mathrm{RD}_k}^n > 0 \tag{5.12}$$

$$T_1 + T_2 = T, \quad T_1 > 0, \quad T_2 > 0 \tag{5.13}$$

and

$$\beta_{\mathrm{RD}_k}^n \in \{0,1\}, \quad \sum_{k}^{K} \beta_{\mathrm{RD}_k}^n = 1 \quad \forall n \in \{1,\ldots,N\} \tag{5.14}$$

It can be deduced that (5.11) is maximized only when $\sum_{m=1}^{N} C_{\mathrm{SR}}^{m} = \sum_{k=1}^{K}\sum_{n=1}^{N} C_{\mathrm{RD}_k}^{n}$. Furthermore, the problem in (5.11) is a combinatorial optimization problem, which is intractable. Therefore, the binary variable $\beta_{\mathrm{RD}_k}^{n}$ is relaxed to a real number within the interval of $(0, 1]$, indicating the time sharing factor of the nth subcarrier between the relay and destination k. Hence, $C_{\mathrm{RD}_k}^{n}$ in (5.10) can be rewritten as

$$C_{\mathrm{RD}_k}^{n} = \frac{T_2 \beta_{\mathrm{RD}_k}^{n}}{T} \log_2 \left(1 + \frac{p_{\mathrm{RD}_k}^{n} \gamma_{\mathrm{RD}_k}^{n}}{\beta_{\mathrm{RD}_k}^{n}}\right) \tag{5.15}$$

Please note that when $\beta_{\mathrm{RD}_k}^{n}$ is close to zero, $C_{\mathrm{RD}_k}^{n}$ in (5.15) also approaches zero.

Thus, (5.11) is modified to

$$\begin{aligned} \underset{p_{\mathrm{SR}}^{m}, p_{\mathrm{RD}_k}^{n}, T_1, T_2, \beta_{\mathrm{RD}_k}^{n}}{\operatorname{argmax}} & \left(\sum_{m=1}^{N} \frac{T_1}{T} \log_2 \left(1 + p_{\mathrm{SR}}^{m} \gamma_{\mathrm{SR}}^{m}\right) \right. \\ & \left. + \sum_{k=1}^{K}\sum_{n=1}^{N} \frac{T_2 \beta_{\mathrm{RD}_k}^{n}}{T} \log_2 \left(1 + \frac{p_{\mathrm{RD}_k}^{n} \gamma_{\mathrm{RD}_k}^{n}}{\beta_{\mathrm{RD}_k}^{n}}\right) \right) \end{aligned} \tag{5.16}$$

subject to (5.12), (5.13), (5.14) and

$$\sum_{m=1}^{N} \frac{T_1}{T} \log_2 \left(1 + p_{\mathrm{SR}}^{m} \gamma_{\mathrm{SR}}^{m}\right) = \sum_{k=1}^{K}\sum_{n=1}^{N} \frac{T_2 \beta_{\mathrm{RD}_k}^{n}}{T} \log_2 \left(1 + \frac{p_{\mathrm{RD}_k}^{n} \gamma_{\mathrm{RD}_k}^{n}}{\beta_{\mathrm{RD}_k}^{n}}\right) \tag{5.17}$$

5.3.2 Optimal Asymmetric Resource Allocation

Since (5.16) is a convex optimization problem, as proved in Appendix C, it can be solved using the KKT conditions [79].

The Lagrangian of the objective function is expressed as

$$\begin{aligned} L \;=\; & \sum_{m=1}^{N} \frac{T_1}{T} \log_2 \left(1 + p_{\mathrm{SR}}^{m} \gamma_{\mathrm{SR}}^{m}\right) + \sum_{k=1}^{K}\sum_{n=1}^{N} \frac{T_2 \beta_{\mathrm{RD}_k}^{n}}{T} \log_2 \left(1 + \frac{p_{\mathrm{RD}_k}^{n} \gamma_{\mathrm{RD}_k}^{n}}{\beta_{\mathrm{RD}_k}^{n}}\right) \\ & - \lambda \left[\sum_{m=1}^{N} \frac{T_1}{T} \log_2 \left(1 + p_{\mathrm{SR}}^{m} \gamma_{\mathrm{SR}}^{m}\right) - \sum_{k=1}^{K}\sum_{n=1}^{N} \frac{T_2 \beta_{\mathrm{RD}_k}^{n}}{T} \log_2 \left(1 + \frac{p_{\mathrm{RD}_k}^{n} \gamma_{\mathrm{RD}_k}^{n}}{\beta_{\mathrm{RD}_k}^{n}}\right) \right] \\ & - \mu \left(\sum_{m=1}^{N} p_{\mathrm{SR}}^{m} + \sum_{k=1}^{K}\sum_{n=1}^{N} p_{\mathrm{RD}_k}^{n} - p_T \right) - \sigma \left(T_1 + T_2 - T\right) \end{aligned} \tag{5.18}$$

where λ, μ and σ are Lagrange multipliers. It can be derived from the KKT condition and constraints (5.12) and (5.13) that $\mu \geq 0$ and $\lambda \in (-1, 1)$.

Letting λ^*, μ^* denote the optimal Lagrange multipliers, it has

$$\begin{aligned}
&\frac{\partial L}{\partial T_1}\Big|_{T_{1,2}=T_{1,2}^*,\beta_{RD_k}^n=\beta_{RD_k}^{*n},p_{SR}^m=p_{SR}^{*m},p_{RD_k}^n=p_{RD_k}^{*n}} \\
&= \frac{1-\lambda^*}{T}\sum_{m=1}^{N}\log_2\left(1+p_{\mathrm{SR}}^{*m}\gamma_{\mathrm{SR}}^m\right)-\sigma^* \\
&\begin{cases} <0 & T_1^*=0 \\ =0 & T_1^*>0 \end{cases}
\end{aligned} \tag{5.19}$$

$$\begin{aligned}
&\frac{\partial L}{\partial T_2}\Big|_{T_{1,2}=T_{1,2}^*,\beta_{RD_k}^n=\beta_{RD_k}^{*n},p_{SR}^m=p_{SR}^{*m},p_{RD_k}^n=p_{RD_k}^{*n}} \\
&= \frac{1+\lambda^*}{T}\sum_{k=1}^{K}\sum_{n=1}^{N}\beta_{RD_k}^{*n}\log_2\left(1+\frac{p_{\mathrm{RD}_k}^{*n}\gamma_{\mathrm{RD}_k}^n}{\beta_{\mathrm{RD}_k}^{*n}}\right)-\sigma^* \\
&\begin{cases} <0 & T_2^*=0 \\ =0 & T_2^*>0 \end{cases}
\end{aligned} \tag{5.20}$$

$$\begin{aligned}
&\frac{\partial L}{\partial p_{SR}^m}\Big|_{T_{1,2}=T_{1,2}^*,\beta_{RD_k}^n=\beta_{RD_k}^{*n},p_{SR}^m=p_{SR}^{*m},p_{RD_k}^n=p_{RD_k}^{*n}} \\
&= \frac{T_1^*(1-\lambda^*)\gamma_{\mathrm{SR}}^m}{T(1+p_{SR}^{*m}\gamma_{\mathrm{SR}}^m)}-\mu^* \\
&\begin{cases} <0 & p_{SR}^{*m}=0 \\ =0 & p_{SR}^{*m}>0 \end{cases}
\end{aligned} \tag{5.21}$$

$$\left.\frac{\partial L}{\partial p^n_{RD_k}}\right|_{T_{1,2}=T^*_{1,2},\beta^n_{RD_k}=\beta^{*n}_{RD_k},p^m_{SR}=p^{*m}_{SR},p^n_{RD_k}=p^{*n}_{RD_k}}$$
$$= \frac{T^*_2(1+\lambda^*)\gamma^n_{\mathrm{RD}_k}\beta^{*n}_{RD_k}}{T(\beta^{*n}_{RD_k}+p^{*n}_{RD_k}\gamma^n_{\mathrm{RD}_k})} - \mu^*$$
$$\begin{cases} <0 & p^{*n}_{RD_k}=0 \\ =0 & p^{*n}_{RD_k}>0 \end{cases} \tag{5.22}$$

and

$$\left.\frac{\partial L}{\partial \beta^n_{RD_k}}\right|_{T_{1,2}=T^*_{1,2},\beta^n_{RD_k}=\beta^{*n}_{RD_k},p^m_{SR}=p^{*m}_{SR},p^n_{RD_k}=p^{*n}_{RD_k}}$$
$$= \frac{T^*_2(1+\lambda^*)}{T}\log_2\left(1+\frac{p^{*n}_{\mathrm{RD}_k}\gamma^n_{\mathrm{RD}_k}}{\beta^{*n}_{\mathrm{RD}_k}}\right) - \frac{T^*_2(1+\lambda^*)p^{*n}_{\mathrm{RD}_k}\gamma^n_{\mathrm{RD}_k}}{T(\beta^{*n}_{\mathrm{RD}_k}+p^{*n}_{\mathrm{RD}_k}\gamma^n_{\mathrm{RD}_k})}$$
$$\begin{cases} <0 & \beta^{*n}_{RD_k}=0 \\ =0 & 0<\beta^{*n}_{RD_k}<1 \\ >0 & \beta^{*n}_{RD_k}=1 \end{cases} \tag{5.23}$$

By using (5.17), (5.19) and (5.20), it is easy to get

$$T_{1,2} = \frac{1\mp\lambda^*}{2}T \tag{5.24}$$

Substituting (5.24) into (5.21) and (5.22), the PA results are

$$p^{*m}_{SR} = \left[\frac{(1-\lambda^*)^2}{2\mu^*} - \frac{1}{\gamma^m_{\mathrm{SR}}}\right]^+ \tag{5.25}$$

and

$$p^{*n}_{\mathrm{RD}_k} = \left[\frac{\beta^{*n}_{\mathrm{RD}_k}(1+\lambda^*)^2}{2\mu^*} - \frac{\beta^{*n}_{\mathrm{RD}_k}}{\gamma^{*n}_{\mathrm{RD}_k}}\right]^+ \tag{5.26}$$

Since $\beta^{*n}_{\mathrm{RD}_k}$ is either 0 or 1, it can be derived from (5.23), (5.24) and (5.26)

that each subcarrier n will be assigned to the user k^* satisfying

$$k^* = \underset{k}{\text{argmax}}(1+\lambda^*)T_2\left(\left[\log_2\left(\frac{(1+\lambda^*)\gamma^n_{\text{RD}_k}T_2}{\mu^*}\right)\right]^+ - \left[\frac{(1+\lambda^*)\gamma^n_{\text{RD}_k}T_2-\mu^*}{(1+\lambda^*)\gamma^n_{\text{RD}_k}T_2}\right]^+\right) \tag{5.27}$$

Hence, the optimal p^m_{SR}, p^n_{RD}, T_1 and T_2 are given by

$$T_1 = \frac{1-\lambda^*}{2}T \tag{5.28}$$

$$T_2 = \frac{1+\lambda^*}{2}T \tag{5.29}$$

$$p^m_{\text{SR}} = \left[\frac{(1-\lambda^*)^2T}{2\mu^*} - \frac{1}{\gamma^m_{\text{SR}}}\right]^+ \tag{5.30}$$

$$p^n_{\text{RD}_k} = \begin{cases} \left[\frac{(1+\lambda^*)^2T}{2\mu^*} - \frac{1}{\gamma^n_{\text{RD}_k}}\right]^+ & k = k^* \\ 0 & k \neq k^* \end{cases} \tag{5.31}$$

The Lagrange multipliers can be obtained by solving the following two equations:

$$\begin{aligned} f(\mu,\lambda) &= p_T - \sum_{m=1}^{N}\left[\frac{(1-\lambda)^2T}{2\mu} - \frac{1}{\gamma^m_{\text{SR}}}\right]^+ \\ &\quad - \sum_{k=1}^{K}\sum_{n=1}^{N}\left[\frac{(1+\lambda)^2T}{2\mu} - \frac{1}{\gamma^n_{\text{RD}_k}}\right]^+ \\ &= 0 \end{aligned} \tag{5.32}$$

$$\begin{aligned} g(\mu,\lambda) &= (1+\lambda)\sum_{k=1}^{K}\sum_{n=1}^{N}\left[\log_2\left(\frac{(1+\lambda)^2T\gamma^n_{\text{RD}_k}}{2\mu}\right)\right]^+ \\ &\quad -(1-\lambda)\sum_{m=1}^{N}\left[\log_2\left(\frac{(1-\lambda)^2T\gamma^m_{\text{SR}}}{2\mu}\right)\right]^+ \\ &= 0 \end{aligned} \tag{5.33}$$

The Bisection algorithm [78] is used to search the optimal Lagrange multipliers μ^* and λ^*. It can be found that $f(\mu, \lambda)$ is a mono-increasing function for $\mu > 0$ and a fixed λ. Define $\gamma_{min} = \min\{\gamma_{\mathrm{SR}}^1, ..., \gamma_{\mathrm{RD}_K}^N\}$, and $\gamma_{max} = \max\{\gamma_{\mathrm{SR}}^1, ..., \gamma_{\mathrm{RD}_K}^N\}$. It can be derived that the lower and upper bounds of the feasible search region of μ are given by $\mu_L = \frac{\gamma_{min} N}{2(p_T \gamma_{min} + N)}$ and $\mu_R = 2T\gamma_{max}$, respectively. Similarly, it can be found that $g(\mu, \lambda)$ is a mono-increasing function with fixed μ. The lower and upper bounds on the feasible search region of λ are $\lambda_L = -1$ and $\lambda_R = 1$, respectively. The search algorithm for the Lagrange multipliers is presented as follows:

1. Initialize λ and μ;
2. Update $\mu_L = \begin{cases} \mu_L & f(\mu, \lambda) > 0 \\ \mu & f(\mu, \lambda) < 0 \end{cases}$, $\mu_R = \begin{cases} \mu_R & f(\mu, \lambda) < 0 \\ \mu & f(\mu, \lambda) > 0 \end{cases}$, and $\mu = (\mu_L + \mu_R)/2$; repeat these procedures until $f(\mu^*, \lambda) = 0$
3. If $g(\mu^*, \lambda) \neq 0$, adjust $\lambda_L = \begin{cases} \lambda_L & g(\mu^*, \lambda) > 0 \\ \lambda & g(\mu^*, \lambda) < 0 \end{cases}$ and $\lambda_R = \begin{cases} \lambda_R & g(\mu^*, \lambda) < 0 \\ \lambda & g(\mu^*, \lambda) > 0 \end{cases}$, and update $\lambda = \frac{\lambda_L + \lambda_R}{2}$;
4. Update μ^* based on Step 2 with the λ obtained in Step 3;
5. Repeat Steps 3 and 4 to obtain μ^* and λ^*, which satisfy both $f(\mu^*, \lambda^*) = 0$ and $g(\mu^*, \lambda^*) = 0$.

Substituting the optimal Lagrange multipliers into (5.27), (5.28), (5.29), (5.30) and (5.31) yields the optimal RA result, including the optimal time duration for each slot, and the optimal power allocated to each subcarrier.

5.3.3 Simulation Results

The advantages of the proposed ARA over the conventional symmetric RA schemes are demonstrated in this subsection. I first compare the required data amount

of feedback information, sensitivity to channel estimation errors and other aspects of the proposed ARA and the conventional symmetric RA schemes in a three-node system, which can be regarded as a special case of the system model with single destination. After that, the comparisons of the system capacity and time allocation are investigated in the multi-destination system as introduced in Section 5.1.

5.3.3.1 Single-Destination Case

Simulation results are provided to show the performance of the proposed ARALF scheme, in comparison with the following symmetric schemes: 1) optimal PA and fixed SP scheme in [86], referred to as sym. time + opt. PA + fixed SP; 2) optimal PA and optimal SP scheme in [43], referred to as sym. time + opt. PA + opt. SP; 3) fixed PA and fixed SP scheme, referred to as sym. time + fixed PA + fixed SP. It assumes perfect channel estimation except in Figure 5.9. Letting $\mathrm{SNR}_{SR} = \frac{p_T E_m\{|h_{SR}^m|^2\}}{2N_0B}$ and $\mathrm{SNR}_{RD} = \frac{p_T E_n\{|h_{RD}^n|^2\}}{2N_0B}$ denote the average SNRs in the two consecutive slots, respectively, $\eta = \mathrm{SNR}_{SR}(\mathrm{dB}) - \mathrm{SNR}_{RD}(\mathrm{dB})$ is defined as the difference between the SNRs for the first slot and the second slot. Normally, the value of η is effected by the distance between the source and the relay and the distance between the relay and the destination. Given a fixed total distance between the source and the destination, the closer the relay to the source, the larger the value of η.

Figure 5.6 demonstrates the normalized amount of feedback data required by the ARA scheme, sym. time + opt. PA + opt. SP [43] and sym. time + opt. PA + fixed SP [86], respectively. With sym. time + opt. PA + opt. SP and sym. time + opt. PA + fixed SP, the CSI of each subcarrier in the second slot needs to be fed back to the source to guarantee $C_{SR}^m = C_{RD}^n$. Hence, the data amount of the feedback information required by sym. time + opt. PA + opt. SP and sym. time + opt. PA + fixed SP is proportional to N, the number of subcarriers.

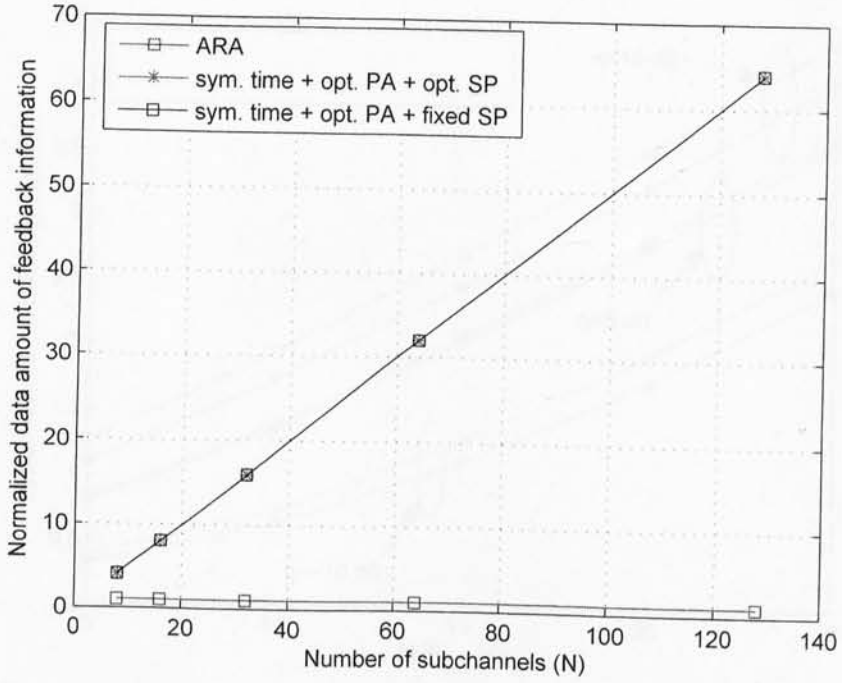

Figure 5.6: Normalized data amount of feedback information

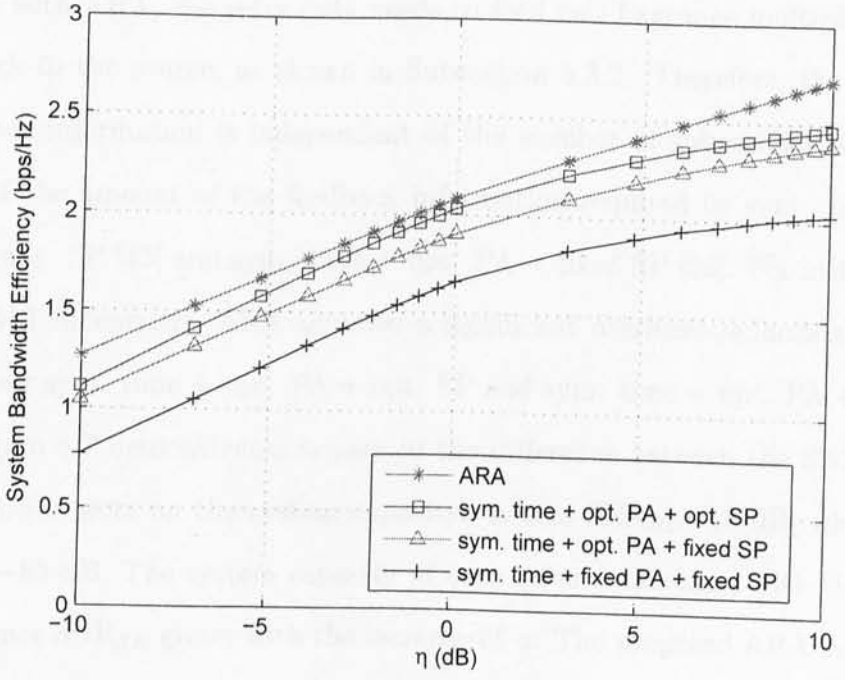

Figure 5.7: Impact of the difference between SNRs for the consecutive slots on the system capacity

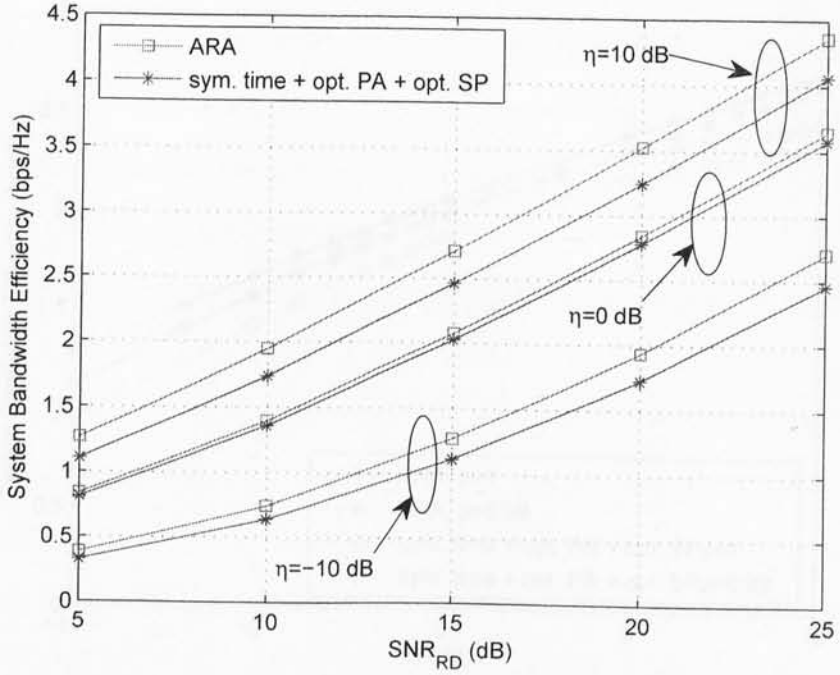

Figure 5.8: Impact of SNR_{RD} on the system capacity

While with ARA, the relay only needs to feed two Lagrange multipliers λ^* and μ^* back to the source, as shown in Subsection 5.3.2. Therefore, the amount of feedback information is independent of the number of subcarriers, and is only $2/N$ of the amount of the feedback information required by sym. time + opt. PA + opt. SP [43] and sym. time + opt. PA + fixed SP [86]. For instance, with $N = 512$ subcarriers, ARA achieves a significant overhead reduction of around 96% over sym. time + opt. PA + opt. SP and sym. time + opt. PA + fixed SP.

Figure 5.7 demonstrates impact of the difference between the SNRs for two consecutive slots on the system capacity. It sets SNR_{RD}=15 dB, which means $N_0 = -85$ dB. The system capacity of each scheme increases with the increase of η, since SNR_{SR} grows with the increase of η. The proposed ARA achieves the highest system capacity, and its performance gains over other symmetric schemes enhance with the increase of $|\eta|$. When $\eta = 10$ dB and $\eta = -10$ dB, the system capacity achieved by ARA is 10% and 15% higher than that achieved by sym.

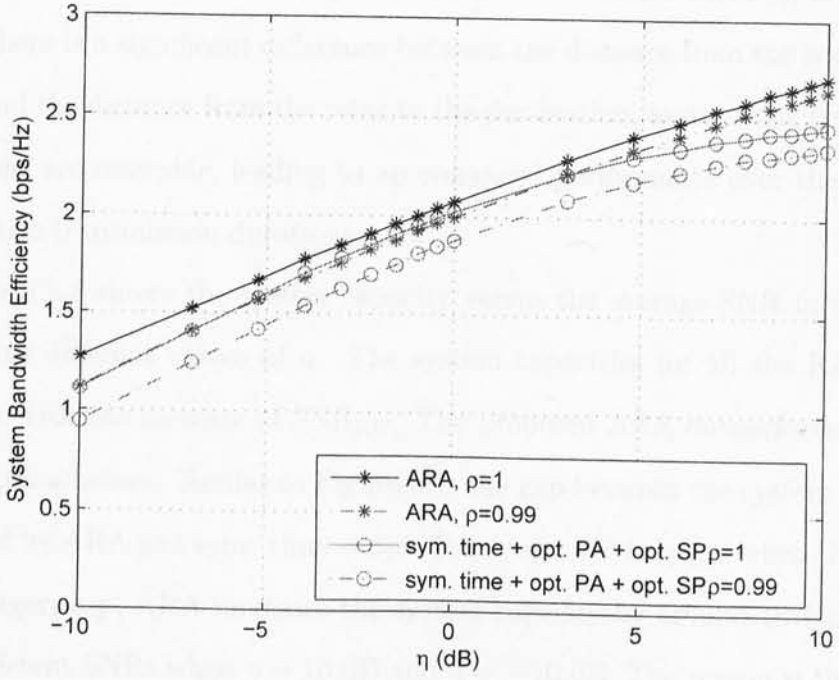

Figure 5.9: Impact of imperfect channel estimation on the system capacity

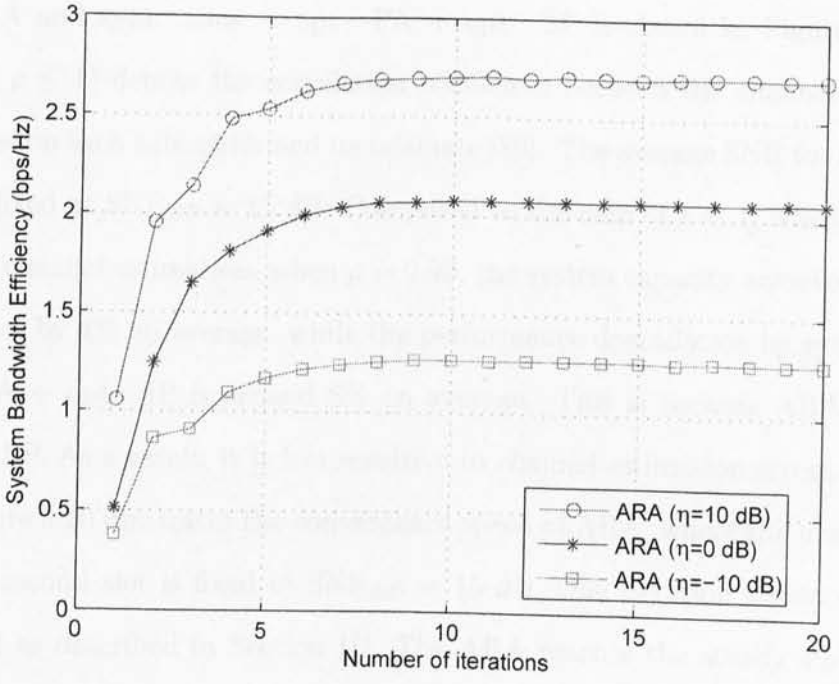

Figure 5.10: Impact of the number of iterations on the system capacity

time + opt. PA + opt. SP, respectively. This is because when $|\eta|$ is large, *i.e.*, when there is a significant difference between the distance from the source to the relay and the distance from the relay to the destination, asymmetric transmission durations are desirable, leading to an enhanced performance over the case with symmetric transmission durations.

Figure 5.8 shows the system capacity versus the average SNR in the second slot, with different values of η. The system capacities for all the RA schemes increase with the increase of SNR_{RD}. The proposed ARA outperforms all other symmetric schemes. Similar to Figure 5.7, the gap between the system capacities achieved by ARA and sym. time + opt. PA + opt. SP is larger when the value of $|\eta|$ is larger, *e.g.*, ARA increases the system capacity by around 10% on average over different SNRs when $\eta = 10$ dB and $\eta = -10$ dB. The reason is the same as that discussed in Figure 5.7.

The impact of imperfect channel estimation on the performances achieved by ARA and sym. time + opt. PA + opt. SP is shown in Figure 5.9. Let ρ ($0 \leq \rho \leq 1$) denote the correlation coefficient between the channel frequency response on each subcarrier and its estimate [89]. The average SNR for the second slot is fixed at $\text{SNR}_{RD} = 15$ dB. Compared to the case of $\rho = 1$, which indicates perfect channel estimation, when $\rho = 0.99$, the system capacity achieved by ARA decreases by 4% on average, while the performance degradation by sym. time + opt. PA + opt. SP is around 8% on average. This is because ARA does not require SP. As a result, it is less sensitive to channel estimation errors.

Figure 5.10 illustrates the convergence speed of ARA, where the average SNR for the second slot is fixed at $\text{SNR}_{RD} = 15$ dB. One iteration consists of Steps 3 and 4 as described in Section III. The ARA reaches the steady state after 8 iterations, demonstrating a fast convergence speed.

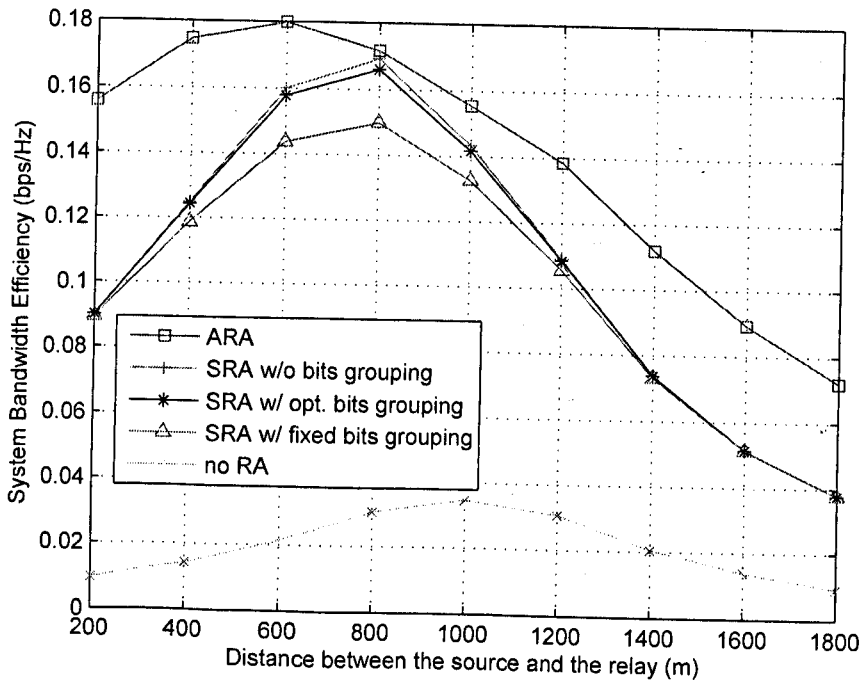

Figure 5.11: Impact of the distance between base station and relay station on system capacity with $K = 32$ mobile users, and a fixed distance of 2000 m between base station and mobile users

5.3.3.2 Multi-Destination Case

The performance of the proposed ARA scheme is demonstrated in this section, in comparison with performances of the following symmetric schemes: 1) optimal symmetric RA scheme in [49], referred to as SRA w/o bits grouping; 2) optimal SA, SP and PA scheme in [46], referred to as SRA w/ opt. bits grouping; 3) optimal SA and PA scheme in [46], referred to as SRA w/ fixed bits grouping; 4) random SA and equal PA scheme, referred to as no RA. A total transmit power of $p_T = 1$ W for the whole relay system is assumed. The total bandwidth is $B = 5$ MHz, which is divided into $N = 512$ subcarriers. The channel has six independent Rayleigh fading paths with an exponentially delay profile and a RMS delay spread of 0.5 μs. It is assumed that the distances between the relay station and mobile users are all equal, *i.e.*, $d_{\mathrm{RD}_1} = ... = d_{\mathrm{RD}_K}$. Let d_{T} denote the total distance between the base station and a mobile user via the relay, *i.e.*,

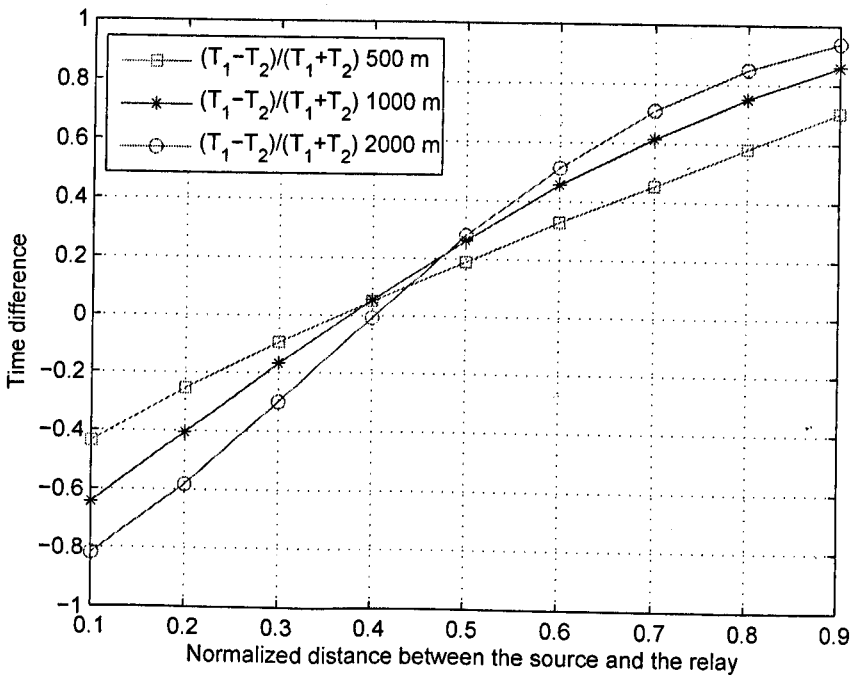

Figure 5.12: Impact of the distance between base station and relay station on time allocation with $K = 32$ mobile users

$d_{\mathrm{T}} = d_{\mathrm{SR}} + d_{\mathrm{RD}_k}$ $(\forall k \in \{1, ..., K\})$. The PL exponent is set to be $\tau = 3.5$ [88], and the PL model is given by $PL(d) = 38.4 + 35 \log_{10}(d)$ dB [88], as a function of distance d in meter (50 m $< d <$ 5000 m). The noise power spectral density is $N_0 = -174$ dBm/Hz.

Figure 5.11 demonstrates the impact of the distance between base station and relay station on the system capacity, with $K = 32$ mobile users and a total distance $d_{\mathrm{T}} = 2000$ m. The proposed asymmetric RA scheme achieves the best performance, and the performance gains over other symmetric schemes enhance with the increase of the difference between the source-to-relay distance and the relay-to-destination distance. The system capacity achieved by ARA is up-to 85% higher (when the source-to-relay distance is 1800 m) than the capacities achieved by other symmetric schemes such as SRA w/o bits grouping [49], SRA w/ opt. bits grouping and SRA w/ fixed bits grouping [46].

Figures 5.12 and 5.13 illustrate the impact of the distance between base station

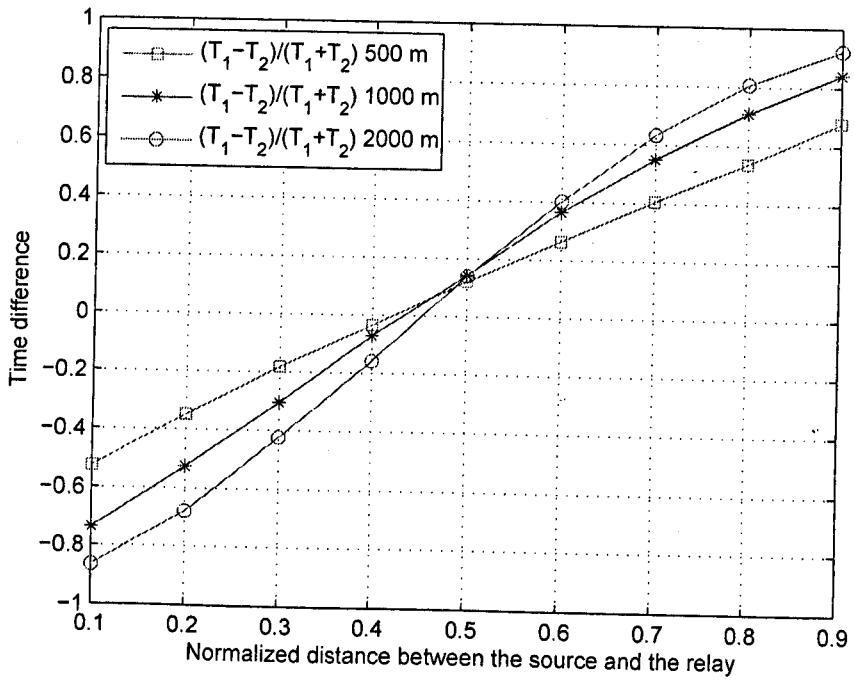

Figure 5.13: Impact of the distance between base station and relay station on time allocation with $K = 4$ mobile users

and relay station on time allocation for two consecutive slots with $K = 32$ and $K = 4$ mobile users, respectively. The time difference increases as total distance augments, especially when the relay is close to the base station or the mobile users, as shown in Figure 5.12. It can also be seen in Figure 5.12 that $T_1 \approx T_2$ when the source-to-relay distance is about 370 m, rather than 500 m, with a total distance $d_T = 1000$ m between the base station and mobile users. This is because that the multiuser diversity increases the average SNR in the second slot. A similar trend can be found in Figure 5.13. It is observed that the transmission durations of the consecutive slots are equal (*i.e.*, $T_1 - T_2 = 0$), when the source-to-relay distance is around 45% of the total distance with $K = 4$ mobile users in Figure 5.13, while it is around 38% with $K = 32$ mobile users in Figure 5.12 due to enhanced multiuser diversity.

In Figure 5.14, the impact of the number of mobile users on time allocation is demonstrated. It is assumed that the average SNR values, $\mathrm{SNR}_{\mathrm{SR}} = \frac{p_T E_m\{|h_{\mathrm{SR}}^m|^2\}}{2N_0 B}$

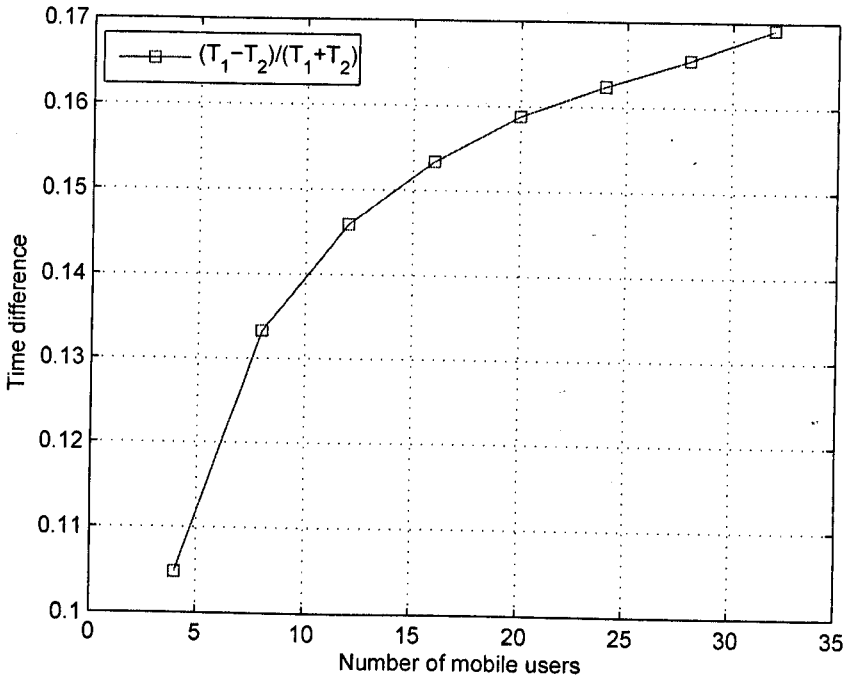

Figure 5.14: Impact of the number of mobile users on time allocation with SNR=15 dB for both of the consecutive slots

and $\mathrm{SNR_{RD}} = \frac{p_T \sum_{k=1}^{K} E_n\left\{|h^n_{\mathrm{RD}_k}|^2\right\}}{2N_0BK}$, for the two consecutive slots are both set to be 15 dB, *e.g.*, the source-to-relay and relay-to-destination distances are equal. The time difference between the two consecutive slots increases with the number of mobile users, which is due to the SNR boost induced by multiuser diversity.

5.4 Summary

The optimal RA scheme for the OFDM based multi-destination regenerative relay system was proposed in Section 5.2, which consists of SA, SP and joint source-relay PA. Simulation results show that among the three steps of the proposed optimal RA, SA has a dominant impact on the system capacity when the relay is closer to the source than to the destinations, due to multiuser diversity, and joint source-relay PA has a larger impact on the system capacity than SP. A

suboptimal RA scheme, which consists of the proposed SA and joint source-relay PA algorithms only, provides a near-optimal performance. Furthermore, this work can be extended to the multi-hop, multi-relay, and multi-destination uplink/downlink OFDM system.

By relaxing the constraint that two subcarriers in consecutive time slots should be paired for data transmission, and designing asymmetric time slots, ARA was proposed in Section 5.3 . Compared to existing symmetric RA schemes, ARA increases the degree of freedom for transmission, and therefore achieves a higher system capacity. The asymmetric time allocation has a significant impact on the system capacity, when the system is with a larger number of mobile users, a longer distance between the base station and mobile users, and when there is a significant difference between the source-to-relay and relay-to-destination distances. ARA is less sensitivity to channel estimation errors than SP based symmetric RA schemes, and contains a fast convergence speed. The simulation results also demonstrate that the two consecutive slot durations become equal when the relay is placed at a point whose distance to the base station is around 40% of the total distance.

Chapter 6

Conclusions and Future Work

6.1 Conclusions

This thesis has investigated the resource management in future wireless communication systems with OFDM and relay technologies, including LTE, WiMAX, LTE-Advanced and WiMAX 2 systems.

In Chapter 4, an adaptive cross-layer design with the PD scheduling at the MAC layer has been proposed, for the downlink multiuser multitasking LTE/WiMAX system with heterogeneous traffic. The proposed suboptimal MWSC based RA algorithm achieves almost the same system performance as the optimal one, at a much lower complexity. Furthermore, it outperforms the MC [3] and PF [38] based RA. The proposed PD scheduling scheme provides significant performance advantages over the queue based M-LWDF [72] and MDU [23] scheduling schemes in terms of the system throughput, BE traffic throughput, QoS traffic delays and QoS traffic outage probabilities, with a wide range of the number of users, and a moderate to high SNR range. It is also proven to achieve the maximum system stability region. By properly choosing the number of packets used for weight calculation, our user based cross-layer design requires a much lower overall complexity than conventional queue based designs.

In Section 5.2, the optimal RA scheme for the OFDM based multi-destination regenerative relay system, which consists of SA, SP and joint source-relay PA, was proposed. Simulation results show that among the three steps of the proposed optimal RA, SA has a dominant impact on the system capacity when the relay is closer to the source than to the destinations, due to multiuser diversity, and joint source-relay PA has a larger impact on the system capacity than SP. A suboptimal RA scheme, which consists of the proposed SA and joint source-relay PA algorithms only, provides a near-optimal performance. Furthermore, this work can be extended to the multi-hop and multi-destination uplink/downlink OFDM system.

Further to proposed the symmetric RA scheme in Section 5.2, an optimal ARA scheme was proposed for relay systems in Section 5.3, by relaxing the constraint that two subcarriers in consecutive time slots should be paired for data transmission, and designing asymmetric time slots. Compared to the existing symmetric RA schemes, ARA increases the degree of freedom for transmission, and therefore achieves a higher system capacity. The asymmetric time allocation has a significant impact on the system capacity, when the system is with a larger number of mobile users, a longer distance between the base station and mobile users, and when there is a significant difference between the source-to-relay and relay-to-destination distances. ARA is less sensitivity to channel estimation errors than SP based symmetric RA schemes, and contains a fast convergence speed. The simulation results also demonstrate that the two consecutive slot durations become equal when the relay is placed at a point whose distance to the base station is around 40% of the total distance.

6.2 Future Work

In the research on relay systems presented in this thesis, it has been assumed that relays are placed far from each other to reduce costs. Under this circumstance, each mobile user could hear from only one of the relays. In practice, some relay coverage may be overlapped, which require the relay selection. Furthermore, to provide various QoSs, it is desirable to introduce cross-layer design in relay systems. Hence, the future research topics are summarized in the following.

- The relay selection will be investigated and incorporated with the proposed symmetric and asymmetric RA schemes, for a relay system where a mobile user can hear more than one relays.

- The proposed RA schemes for relay systems have only considered maximizing the overall system capacity, without considering the fairness among mobile users. This will be addressed in the future.

- Cross-layer designs will be introduced into the relay systems, with various design criteria including delay bound, outage probability, BER and etc. For example, RA at the PHY layer and scheduling at the MAC layer will be performed via cross-talk at the same time.

Appendix A

Proof of Theorem 4.1

Define $\delta_k = \lambda_k - R_k$ ($\forall k \in \{1, ..., K\}$) as the difference between the data arrival rate λ_k and the data transmission rate R_k for user k in current slot L, which is a finite value since both λ_k and R_k are finite. When the average data arrival rates locate in the interior of the capacity region $\widetilde{C}$, there exists at least one average transmission rate vector $E\{\mathbf{R}\} \in \widetilde{C}$ so that $E\{\lambda_k\} < E\{R_k\}$ for all $k \in \{1, ..., K\}$, *i.e.*, $E\{\delta_k\} < 0$.

A Lyapunov function [82] $Y(\mathbf{Q}_Z) = \sum_{k=1}^{K} y(Q_{k,Z})$ is defined, where $y(Q_{k,Z})$ is a nonnegative function and its derivative is given by

$$g(Q_{k,Z}) = \frac{\partial y(Q_{k,Z})}{\partial Q_{k,Z}} = \beta_{\max} Q_{k,Z} \tag{A.1}$$

where $\beta_{\max} = \underset{k,i}{\operatorname{argmax}} \beta_{k,i}$. It is obvious that $g(x) \in [0, \infty)$ ($\forall x < \infty$).

Define $\Delta_{k,Z} = Q_{k,Z+1} - Q_{k,Z}$ as the difference between the queue lengths in two consecutive slots $(Z+1)$ and Z for user k. Using the mean value theorem

[90], the Lyapunov drift can be written as:

$$\begin{aligned} & E\{Y(\mathbf{Q}_{Z+1}) - Y(\mathbf{Q}_Z)\} \\ &= \sum_{k=1}^{K} E\{y(Q_{k,Z+1}) - y(Q_{k,Z})\} \\ &= \sum_{k=1}^{K} E\{\Delta_{k,Z} g(Q_{k,Z} + v_{k,Z}\Delta_{k,Z})\} \end{aligned} \tag{A.2}$$

where $0 < v_{k,Z} < 1$ ($\forall k \in \{1, ..., K\}$). According to (A.1), it is straightforward to have

$$\begin{aligned} & \Delta_{k,Z} g(Q_{k,Z} + v_{k,Z}\Delta_{k,Z}) \\ &= \Delta_{k,Z}\beta_{\max} Q_{k,Z} + \Delta_{k,Z}^2 \beta_{\max} v_{k,Z} \end{aligned} \tag{A.3}$$

Two scenarios for (A.3) are discussed in the following. Note that the data arriving in slot Z will be served in subsequent slot(s).

Scenario 1: $Q_{k,Z} \leq R_k T_{\text{slot}}$, *i.e.*, all the to-be-served data of user k can be sent out during slot Z. Meanwhile, the amount of data that arrives in slot Z and is to be served in subsequent slot(s) is $\lambda_k T_{\text{slot}}$. Hence, the queue length for user k at the beginning of slot $(Z+1)$ is $Q_{k,Z+1} = \lambda_k T_{\text{slot}}$, and the difference between the queue lengths of slots $(Z+1)$ and Z for user k is $\Delta_{k,Z} = Q_{k,Z+1} - Q_{k,Z} = \lambda_k T_{\text{slot}} - Q_{k,Z}$. Since $0 \leq Q_{k,Z} \leq R_k T_{\text{slot}}$, it has $\delta_k T_{\text{slot}} \leq \Delta_{k,Z} \leq \lambda_k T_{\text{slot}}$. Since δ_k and λ_k are finite values, it is easy to derive that there exists an upper bound $\sigma_1 \in (0, \infty)$ for (A.3), *i.e.*, $\Delta_{k,Z} g(Q_{k,Z} + v_{k,Z}\Delta_{k,Z}) \leq \sigma_1$.

Scenario 2: $Q_{k,Z} > R_k T_{\text{slot}}$, *i.e.*, not all the to-be-served data of user k can be sent out during slot Z. Hence, the data of size $(Q_{k,Z} - R_k T_{\text{slot}})$ is left unserved at the end of slot Z. Meanwhile, the data of size $\lambda_k T_{\text{slot}}$ arrives in slot Z, waiting to be served in subsequent slot(s). Thus, the to-be-served queue length for user k at the beginning of slot $(Z+1)$ is $Q_{k,Z+1} = Q_{k,Z} - R_k T_{\text{slot}} + \lambda_k T_{\text{slot}} = Q_{k,Z} + \delta_k T_{\text{slot}}$.

Thus, the difference between queue lengths of slots $(Z+1)$ and Z for user k is $\Delta_{k,Z} = Q_{k,Z+1} - Q_{k,Z} = \delta_k T_{\text{slot}}$. In this case, it is easy to derive that there exists an upper bound $\sigma_2 \in (0, \infty)$ for $\Delta_{k,Z}^2 \beta_{\max} v_{k,Z}$ in (A.3), *i.e.*, $\Delta_{k,Z}^2 \beta_{\max} v_{k,Z} \leq \sigma_2$. Therefore, it has $\Delta_{k,Z} g(Q_{k,Z} + v_{k,Z}\Delta_{k,Z}) \leq \sigma_2 + \delta_k T_{\text{slot}} \beta_{\max} Q_{k,Z}$.

Using the above discussion for two scenarios, it can be obtained that:

$$\begin{aligned} & E\{\Delta_{k,Z} g(Q_{k,Z} + v_{k,Z}\Delta_{k,Z})\} \\ & \leq \sigma_1 + \sigma_2 + E\{\delta_k T_{\text{slot}} \beta_{\max} Q_{k,Z}\} \\ & \leq \sigma_1 + \sigma_2 + \overline{\delta}_{\max} T_{\text{slot}} E\{\beta_{\max} Q_{k,Z}\} \end{aligned} \tag{A.4}$$

where $\overline{\delta}_{\max} = \underset{k}{\text{argmax}}\,(E\{\delta_k\}) < 0$. From (4.25) and (4.28), it can be derived that

$$\beta_{\max} Q_{k,Z} \geq \sum_{i=1}^{I_k} \sum_{l \in \mathbb{L}_{k,i}^Z} W_{k,i,l} \geq 0 \tag{A.5}$$

Substituting (A.5) into (A.4) leads to

$$\begin{aligned} & E\{\Delta_{k,Z} g(Q_{k,Z} + v_{k,Z}\Delta_{k,Z})\} \\ & \leq \sigma_1 + \sigma_2 + \overline{\delta}_{\max} T_{\text{slot}} \sum_{i=1}^{I_k} \sum_{l \in \mathbb{L}_{k,i}^Z} E\{W_{k,i,l}\} \end{aligned} \tag{A.6}$$

From (A.2) and (A.6), it is easy to have

$$\begin{aligned} & E\{Y(\mathbf{Q}_{Z+1}) - Y(\mathbf{Q}_Z)\} \\ & \leq K(\sigma_1 + \sigma_2) + \overline{\delta}_{\max} T_{\text{slot}} \sum_{k=1}^{K} \sum_{i=1}^{I_k} \sum_{l \in \mathbb{L}_{k,i}^Z} E\{W_{k,i,l}\} \end{aligned} \tag{A.7}$$

Letting $a = K(\sigma_1 + \sigma_2) > 0$ and $b = -\overline{\delta}_{\max} T_{\text{slot}} > 0$, (A.7) reduces to

$$\begin{aligned} & E\left\{Y(\mathbf{Q}_{Z+1}) - Y(\mathbf{Q}_Z)\right\} \\ & \leq a - b \sum_{k=1}^{K} \sum_{i=1}^{I_k} \sum_{l \in \mathbb{L}_{k,i}^Z} E\left\{W_{k,i,l}\right\} \quad \forall Z \in [1, \infty) \end{aligned} \tag{A.8}$$

Using (A.8), the Lyapunov drift $E\{Y(\mathbf{Q}_{Z+1}) - Y(\mathbf{Q}_Z)\}$ averaged over slots $Z = 1$ to P is given by

$$\begin{aligned} & \frac{1}{P} \sum_{Z=1}^{P} E\left\{Y(\mathbf{Q}_{Z+1}) - Y(\mathbf{Q}_Z)\right\} \\ & = \frac{1}{P}\left[E\left\{Y(\mathbf{Q}_{P+1}) - Y(\mathbf{Q}_1)\right\}\right] \\ & \leq a - \frac{b}{P} \sum_{k=1}^{K} \sum_{i=1}^{I_k} \sum_{Z=1}^{P} \sum_{l \in \mathbb{L}_{k,i}^Z} E\left\{W_{k,i,l}\right\} \end{aligned} \tag{A.9}$$

Since $Y(\mathbf{Q}_Z) \geq 0$ $(\forall Z \in [1, \infty))$, it can be derived from (A.9) that

$$\begin{aligned} & \frac{1}{P} \sum_{k=1}^{K} \sum_{i=1}^{I_k} \sum_{Z=1}^{P} \sum_{l \in \mathbb{L}_{k,i}^Z} E\left\{W_{k,i,l}\right\} \\ & \leq \frac{a}{b} + \frac{1}{bP}\left[E\left\{Y(\mathbf{Q}_1) - Y(\mathbf{Q}_{P+1})\right\}\right] \\ & \leq \frac{a}{b} + \frac{E\left\{Y(\mathbf{Q}_1)\right\}}{bP} \end{aligned} \tag{A.10}$$

Letting $P \to \infty$ and taking lim sup of (A.10) yield

$$\limsup_{P \to \infty} \frac{1}{P} \sum_{k=1}^{K} \sum_{i=1}^{I_k} \sum_{Z=1}^{P} \sum_{l \in \mathbb{L}_{k,i}^Z} E\left\{W_{k,i,l}\right\} \leq \frac{a}{b} \tag{A.11}$$

According to (4.25), it can be obtained obtained that

$$\begin{aligned}
&\limsup_{P\to\infty}\frac{1}{P}\sum_{k=1}^{K}\sum_{i=1}^{I_k}\sum_{Z=1}^{P}\sum_{l\in\mathbb{L}_{k,i}^{Z}}E\{D_{k,i,l}\} \qquad (A.12)\\
&=\limsup_{P\to\infty}\frac{1}{P}\sum_{k=1}^{K}\sum_{i=1}^{I_k}\sum_{Z=1}^{P}\sum_{l\in\mathbb{L}_{k,i}^{Z(U)}}E\left\{\frac{W_{k,i,l}}{\beta_{k,i}}\right\}\\
&+\limsup_{P\to\infty}\frac{1}{P}\sum_{k=1}^{K}\sum_{i=1}^{I_k}\sum_{Z=1}^{P}\sum_{l\in\overline{\mathbb{L}_{k,i}^{Z(U)}}}E\left\{\frac{(C_{k,i,l}+1)W_{k,i,l}}{\beta_{k,i}}\right\}\\
&=\limsup_{P\to\infty}\frac{1}{P}\sum_{k=1}^{K}\sum_{i=1}^{I_k}\sum_{Z=1}^{P}\sum_{l\in\mathbb{L}_{k,i}^{Z}}E\left\{\frac{W_{k,i,l}}{\beta_{k,i}}\right\}\\
&+\limsup_{P\to\infty}\frac{1}{P}\sum_{k=1}^{K}\sum_{i=1}^{I_k}\sum_{Z=1}^{P}\sum_{l\in\overline{\mathbb{L}_{k,i}^{Z(U)}}}E\left\{\frac{C_{k,i,l}W_{k,i,l}}{\beta_{k,i}}\right\}
\end{aligned}$$

where $\mathbb{L}_{k,i}^{Z(U)}$ denotes the set of packets belonging to queue i of user k and falling in the guard interval at slot Z, and $\overline{\mathbb{L}_{k,i}^{Z(U)}}$ denotes the set consisting of the rest packets. Since $W_{k,i,l}\in[0,\infty)$ and $\beta_{k,i}\in[1,\infty)$, the first term of (A.12) satisfies

$$\begin{aligned}
&\limsup_{P\to\infty}\frac{1}{P}\sum_{k=1}^{K}\sum_{i=1}^{I_k}\sum_{Z=1}^{P}\sum_{l\in\mathbb{L}_{k,i}^{Z}}E\left\{\frac{W_{k,i,t}}{\beta_{k,i}}\right\}\\
&<\limsup_{P\to\infty}\frac{1}{P}\sum_{k=1}^{K}\sum_{i=1}^{I_k}\sum_{Z=1}^{P}\sum_{l\in\mathbb{L}_{k,i}^{Z}}E\{W_{k,i,l}\} \qquad (A.13)
\end{aligned}$$

In the second term of (A.12), $C_{k,i,l}$ is in the range of $0<C_{k,i,l}\le U_{k,i}-G_{k,i}$. Thus, there exists $M\in[0,\infty)$ so that $C_{k,i,l}W_{k,i,l}/\beta_{k,i}<M$. Letting $U_{\max}=\underset{k,i}{\operatorname{argmax}}U_{k,i}$ and $I_{\max}=\underset{k}{\operatorname{argmax}}I_k$, the second term of (A.12) satisfies

$$\begin{aligned}
&\limsup_{P\to\infty}\frac{1}{P}\sum_{k=1}^{K}\sum_{i=1}^{I_k}\sum_{Z=1}^{P}\sum_{l\in\overline{\mathbb{L}_{k,i}^{Z(U)}}}E\left\{\frac{C_{k,i,l}W_{k,i,l}}{\beta_{k,i}}\right\}\\
&<KI_{\max}M(U_{k,i}-G_{k,i})<KI_{\max}MU_{\max} \qquad (A.14)
\end{aligned}$$

Using (A.13) and (A.14), (A.12) can be written as

$$\limsup_{P\to\infty} \frac{1}{P} \sum_{k=1}^{K} \sum_{i=1}^{I_k} \sum_{Z=1}^{P} \sum_{l\in\mathbb{L}_{k,i}^{Z}} E\{D_{k,i,l}\} \tag{A.15}$$

$$< \limsup_{P\to\infty} \frac{1}{P} \sum_{k=1}^{K} \sum_{i=1}^{I_k} \sum_{Z=1}^{P} \sum_{l\in\mathbb{L}_{k,i}^{Z}} E\{W_{k,i,l}\} + KI_{\max}MU_{\max}$$

Substituting (A.11) into (A.15), and using $a, b \in (0, \infty)$ yield

$$\limsup_{P\to\infty} \frac{1}{P} \sum_{k=1}^{K} \sum_{i=1}^{I_k} \sum_{Z=1}^{P} \sum_{l\in\mathbb{L}_{k,i}^{Z}} E\{D_{k,i,l}\}$$

$$< \frac{a}{b} + KI_{\max}MU_{\max} < \infty \tag{A.16}$$

Therefore, the system is stable as defined by (4.30).

Appendix B

Proof of Theorem 5.1

It is hereby to prove that no other RA results, whether they are obtained via joint or in-serial RA, will achieve a higher system capacity than the proposed 3-step RA scheme does. In the Appendix, α, β and θ are used to represent the proposed SA, SP and PA algorithms, respectively, and the resulting capacity is denoted by $C(\alpha, \beta, \theta)$. For any joint RA algorithm, there exists an equivalent set of in-serial algorithms α', β' and θ', which produce the same results as the joint RA scheme. The corresponding capacity is denoted by $C(\alpha', \beta', \theta')$. In the following, it shows that

$$C(\alpha, \beta, \theta) \geq C(\alpha', \beta', \theta') \tag{B.1}$$

Scenario 1: It first proves

$$C(\alpha', \beta', \theta) > C(\alpha', \beta', \theta') \tag{B.2}$$

where $\theta' \neq \theta$. In order to prove (B.2), an intermediate PA algorithm $\dot{\theta}$ is introduced, which is obtained by partly changing the PA algorithm θ, and prove $C(\alpha', \beta', \theta) > C(\alpha', \beta', \dot{\theta}) > C(\alpha', \beta', \theta')$. Assume $\dot{\theta}$ is the same as θ, except for two subcarrier pairs $\{m, n\}_i$ and $\{m', n'\}_j$, whose resulting aggregate powers are respectively $(p_i^{m,n} + \Delta p)$ and $(p_j^{m',n'} - \Delta p)$, where $p_i^{m,n}$ and $p_j^{m',n'}$ are the allocated

powers by θ, and $\Delta p > 0$. The difference between the two system capacities achieved by (α', β', θ) and (α', β', $\dot{\theta}$) is given by:

$$\begin{aligned}
&C(\alpha', \beta', \theta) - C(\alpha', \beta', \dot{\theta}) \\
&= \frac{1}{2}\log_2\left[\left(1 + a_i^{m,n} p_i^{m,n}\right)\left(1 + a_j^{m',n'} p_j^{m',n'}\right)\right] \\
&-\frac{1}{2}\log_2\left\{\left[1 + a_i^{m,n}\left(p_i^{m,n} + \Delta p\right)\right]\left[1 + a_j^{m',n'}\left(p_j^{m',n'} - \Delta p\right)\right]\right\}
\end{aligned} \tag{B.3}$$

Since the aggregate powers $p_i^{m,n}$ and $p_j^{m',n'}$ are deduced from the water-filling algorithm [39], it has

$$a_j^{m',n'}\left(1 + a_i^{m,n} p_i^{m,n}\right) = a_i^{m,n}\left(1 + a_j^{m',n'} p_j^{m',n'}\right) \tag{B.4}$$

Using (B.4) and $\Delta p > 0$, it can be derived

$$\begin{aligned}
&\left(1 + a_i^{m,n} p_i^{m,n}\right)\left(1 + a_j^{m',n'} p_j^{m',n'}\right) \\
&-\left[1 + a_i^{m,n}\left(p_i^{m,n} + \Delta p\right)\right]\left[1 + a_j^{m',n'}\left(p_j^{m',n'} - \Delta p\right)\right] \\
&= a_j^{m',n'}\left(1 + a_i^{m,n} p_i^{m,n}\right)\Delta p - a_i^{m,n}\left(1 + a_j^{m',n'} p_j^{m',n'}\right)\Delta p + a_i^{m,n} a_j^{m',n'}\Delta p^2 \\
&= a_i^{m,n} a_j^{m',n'}\Delta p^2 > 0
\end{aligned} \tag{B.5}$$

Using (B.5), it is straightforward to obtain

$$C(\alpha', \beta', \theta) > C(\alpha', \beta', \dot{\theta}) \tag{B.6}$$

Since PA algorithm θ' can be obtained by partly changing the PA algorithm $\dot{\theta}$, and it can be derived from (B.6) that there exists an intermediate PA algorithm $\dot{\theta}$ satisfying

$$C(\alpha', \beta', \theta) > C(\alpha', \beta', \dot{\theta}) > C(\alpha', \beta', \theta') \tag{B.7}$$

Scenario 2: It has been proved in *Scenario 1* that $C(\alpha', \beta', \theta) > C(\alpha', \beta', \theta')$

$(\theta' \neq \theta)$. If $\beta' \neq \beta$, it is to prove

$$C(\alpha', \beta, \theta) > C(\alpha', \beta', \theta) \tag{B.8}$$

Consider two RA schemes $(\alpha', \dot{\beta}, \theta)$ and (α', β, θ), where $\dot{\beta}$ denotes an SP algorithm different from β. Assume the two RA schemes achieve the same SP result except that two subcarrier pairs are exchanged, *i.e.*, $(m,n)_i$ and $(m',n')_j$, achieved by (α', β, θ), whose aggregate powers are respectively $p_i^{m,n}$ and $p_j^{m',n'}$, are changed to $(m',n)_i$ and $(m,n')_j$, whose powers are $p_i^{m',n}$ and $p_j^{m,n'}$, respectively. Also assume $\gamma_{SR}^{m} > \gamma_{SR}^{m'}$ and $\gamma_{RD_i}^{n} > \gamma_{RD_j}^{n'}$. In the following, it is to prove $C(\alpha', \beta, \theta) > C(\alpha', \dot{\beta}, \theta)$. Since it is hard to compare the two system capacities directly, an intermediate system capacity $C(\alpha', \beta, \dot{\theta})$ is introduced, where $\dot{\theta} \neq \theta$. Since the optimality of the proposed PA has been proved in *Scenario 1*, it shows $C(\alpha', \beta, \theta) > C(\alpha', \beta, \dot{\theta})$. It now proves $C(\alpha', \beta, \dot{\theta}) > C(\alpha', \dot{\beta}, \theta)$ by considering the following two cases.

1) If $a_i^{m',n} \geq a_j^{m,n'}$: Assume the PA result of $(\alpha', \beta, \dot{\theta})$ is the same as that of $(\alpha', \dot{\beta}, \theta)$. The difference between capacities $C(\alpha', \beta, \dot{\theta})$ and $C(\alpha', \dot{\beta}, \theta)$ is given by

$$\begin{aligned} C(\alpha', \beta, \dot{\theta}) - C(\alpha', \dot{\beta}, \theta) = & \frac{1}{2}\log_2\left[\left(1 + a_i^{m,n} p_i^{m',n}\right)\left(1 + a_j^{m',n'} p_j^{m,n'}\right)\right] \\ & - \frac{1}{2}\log_2\left[\left(1 + a_i^{m',n} p_i^{m',n}\right)\left(1 + a_j^{m,n'} p_j^{m,n'}\right)\right] \end{aligned} \tag{B.9}$$

Define $A = \frac{1}{a_i^{m',n}} + \frac{1}{a_j^{m,n'}} = \frac{1}{a_i^{m,n}} + \frac{1}{a_j^{m',n'}}$, $x = \frac{1}{a_j^{m',n'}} - \frac{1}{a_i^{m,n}}$ and $x' = \frac{1}{a_j^{m,n'}} - \frac{1}{a_i^{m',n}}$. Since $\gamma_{SR}^{m} > \gamma_{SR}^{m'}$, $\gamma_{RD_i}^{n} > \gamma_{RD_j}^{n'}$, and $a_i^{m',n} \geq a_j^{m,n'}$, it is easy to prove $p_i^{m',n} \geq p_j^{m,n'}$ according to *Proposition 3*, and to prove $0 \leq x' < x < A$. It can be derived

$a_i^{m,n} = \frac{2}{A-x}$, $a_j^{m',n'} = \frac{2}{A+x}$, $a_i^{m',n} = \frac{2}{A-x'}$ and $a_j^{m,n'} = \frac{2}{A+x'}$. Hence, it has

$$\begin{aligned}
&\left(1 + a_i^{m,n} p_i^{m',n}\right)\left(1 + a_j^{m',n'} p_j^{m,n'}\right) - \left(1 + a_i^{m',n} p_i^{m',n}\right)\left(1 + a_j^{m,n'} p_j^{m,n'}\right) \\
&= \frac{2p_i^{m',n}(A+x) + 2p_j^{m,n'}(A-x) + 4p_i^{m',n} p_j^{m,n'}}{A^2 - x^2} \\
&\quad - \frac{2p_i^{m',n}(A+x') + 2p_j^{m,n'}(A-x') + 4p_i^{m',n} p_j^{m,n'}}{A^2 - x'^2} \\
&> \frac{2\left(p_i^{m',n} - p_j^{m,n'}\right)(x - x')}{A^2 - x'^2} > 0
\end{aligned} \tag{B.10}$$

Using (B.9) and (B.10), it can be deduced that $C(\alpha', \beta, \dot{\theta}) > C(\alpha', \dot{\beta}, \theta)$.

2) If $a_i^{m',n} < a_j^{m,n'}$: Assume the PA result of $\dot{\theta}$ is the same as θ, except for two subcahnnel pairs $\{m, n\}_i$ and $\{m', n'\}_j$, whose aggregate powers are equal to $p_j^{m,n'}$ and $p_i^{m',n}$, respectively. The difference between capacities $C(\alpha', \beta, \dot{\theta})$ and $C(\alpha', \dot{\beta}, \theta)$ is given by

$$\begin{aligned}
C(\alpha', \beta, \dot{\theta}) - C(\alpha', \dot{\beta}, \theta) &= \frac{1}{2}\log_2\left[\left(1 + a_i^{m,n} p_j^{m,n'}\right)\left(1 + a_j^{m',n'} p_i^{m',n}\right)\right] \\
&\quad - \frac{1}{2}\log_2\left[\left(1 + a_i^{m',n} p_i^{m',n}\right)\left(1 + a_j^{m,n'} p_j^{m,n'}\right)\right]
\end{aligned} \tag{B.11}$$

Since $\gamma_{SR}^{m} > \gamma_{SR}^{m'}$, $\gamma_{RD_i}^{n} > \gamma_{RD_j}^{n'}$ and $a_i^{m',n} < a_j^{m,n'}$, it proves $p_i^{m',n} < p_j^{m,n'}$ according to *Proposition 3*, and to prove $0 \le -x' < x < A$. Hence, it is straightforward to have

$$\begin{aligned}
&\left(1 + a_i^{m,n} p_j^{m,n'}\right)\left(1 + a_j^{m',n'} p_i^{m',n}\right) - \left(1 + a_i^{m',n} p_i^{m',n}\right)\left(1 + a_j^{m,n'} p_j^{m,n'}\right) \\
&> \frac{2\left(p_j^{m,n'} - p_i^{m',n}\right)(x + x')}{A^2 - x'^2} > 0
\end{aligned} \tag{B.12}$$

Using (B.11) and (B.12), it can be deduced that $C(\alpha', \beta, \dot{\theta}) > C(\alpha', \dot{\beta}, \theta)$.

Since $C(\alpha', \beta, \theta) > C(\alpha', \beta, \dot{\theta})$ and $C(\alpha', \beta, \dot{\theta}) > C(\alpha', \dot{\beta}, \theta)$, it can be obtained that

$$C(\alpha', \beta, \theta) > C(\alpha', \dot{\beta}, \theta) \tag{B.13}$$

SP algorithm β' can be obtained by applying a step-wise exchange algorithm [41] to β, where two selected neighboring subcarriers in the second slot are exchanged in each step, and the system capacity is reduced according to (B.13). Therefore, it can be derived that

$$C(\alpha', \beta, \theta) > C(\alpha', \dot{\beta}, \theta) > C(\alpha', \beta', \theta) \tag{B.14}$$

Scenario 3: It can be deduced from *Scenario 1* and *Scenario 2* that

$$C(\alpha', \beta, \theta) > C(\alpha', \beta', \theta) > C(\alpha', \beta', \theta') \tag{B.15}$$

where $\beta' \neq \beta$ and $\theta' \neq \theta$. The optimality of the proposed RA scheme is shown by proving

$$C(\alpha, \beta, \theta) > C(\alpha', \beta, \theta) \tag{B.16}$$

where $\alpha' \neq \alpha$.

Define $\dot{\alpha}$ as an intermediate SA algorithm, which obtains the same SA result as α, except that a subcarrier pair $(m,n)_i$ in α, where $\gamma^n_{RD_i} > \gamma^n_{RD_j}$ $(\forall i,j \in \{1,...,K\}$, is re-allocated to destination j by $\dot{\alpha}$. In the following, it is to prove $C(\alpha, \beta, \theta) > C(\dot{\alpha}, \beta, \theta)$.

Since it is hard to compare the two system capacities $C(\alpha, \beta, \theta)$ and $C(\dot{\alpha}, \beta, \theta)$ directly, an intermediate system capacity $C(\alpha, \beta, \dot{\theta})$ is introduced, and it is to prove $C(\alpha, \beta, \theta) > C(\alpha, \beta, \dot{\theta}) > C(\dot{\alpha}, \beta, \theta)$, where $\dot{\theta}$ provides the same PA result as that of $(\dot{\alpha}, \beta, \theta)$. Due to the optimality of θ, which is shown in *Scenario 1*, it have $C(\alpha, \beta, \theta) > C(\alpha, \beta, \dot{\theta})$. It is to prove $C(\alpha, \beta, \dot{\theta}) > C(\dot{\alpha}, \beta, \theta)$. Letting $p_j^{m,n}$ denote the aggregate power for $(m,n)_j$, the difference between the two system capacities is given by

$$C(\alpha, \beta, \dot{\theta}) - C(\dot{\alpha}, \beta, \theta) = \frac{1}{2}\log_2\left(1 + a_i^{m,n} p_j^{m,n}\right) - \frac{1}{2}\log_2\left(1 + a_j^{m,n} p_j^{m,n}\right) \tag{B.17}$$

Since $a_i^{m,n} > a_j^{m,n}$, $C(\alpha,\beta,\dot{\theta}) - C(\dot{\alpha},\beta,\theta) > 0$ in (B.17). By using $C(\alpha,\beta,\theta) > C(\alpha,\beta,\dot{\theta})$ and $C(\alpha,\beta,\dot{\theta}) > C(\dot{\alpha},\beta,\theta)$, it can be obtained that

$$C(\alpha,\beta,\theta) > C(\dot{\alpha},\beta,\theta) \tag{B.18}$$

Similar to *Scenario 2*, SA algorithm α' can be obtained by applying a step-wise exchange algorithm [41] to α, and the system capacity is reduced in each step according to (B.18). Therefore, there exists an intermediate SA algorithm $\dot{\alpha}$ satisfying that $C(\alpha,\beta,\theta) > C(\dot{\alpha},\beta,\theta) > C(\alpha',\beta,\theta)$.

It can be deduced from the above discussion that

$$C(\alpha,\beta,\theta) > C(\alpha',\beta,\theta) > C(\alpha',\beta',\theta) > C(\alpha',\beta',\theta') \tag{B.19}$$

where $\alpha' \neq \alpha$, $\beta' \neq \beta$ and $\theta' \neq \theta$. It means that the proposed RA scheme achieves a higher system capacity than any other RA schemes. Therefore, the proposed RA scheme is optimal.

Appendix C

Proof of Convexity of (5.16)

Using (5.13) and (5.17), T_1 and T_2 are given by

$$T_1 = \frac{\sum_k \sum_n \beta_k^n \log_2\left(1+\frac{p_{\mathrm{RD}}^n \gamma_{\mathrm{RD}}^n}{\beta_k^n}\right) T}{\sum_m \log_2\left(1+p_{\mathrm{SR}}^m \gamma_{\mathrm{SR}}^m\right) + \sum_k \sum_n \beta_k^n \log_2\left(1+\frac{p_{\mathrm{RD}}^n \gamma_{\mathrm{RD}}^n}{\beta_k^n}\right)} \tag{C.1}$$

$$T_2 = \frac{\sum_m \log_2\left(1+p_{\mathrm{SR}}^m \gamma_{\mathrm{SR}}^m\right) T}{\sum_m \log_2\left(1+p_{\mathrm{SR}}^m \gamma_{\mathrm{SR}}^m\right) + \sum_k \sum_n \beta_k^n \log_2\left(1+\frac{p_{\mathrm{RD}}^n \gamma_{\mathrm{RD}}^n}{\beta_k^n}\right)} \tag{C.2}$$

Substituting (C.1) and (C.2) into (5.16), the objective function of ARA becomes

$$\underset{p_{\mathrm{SR}}^m, p_{\mathrm{RD}}^n}{\operatorname{argmax}} \frac{2\sum_m \log_2\left(1+p_{\mathrm{SR}}^m \gamma_{\mathrm{SR}}^m\right) \sum_k \sum_n \beta_k^n \log_2\left(1+\frac{p_{\mathrm{RD}}^n \gamma_{\mathrm{RD}}^n}{\beta_k^n}\right)}{\sum_m \log_2\left(1+p_{\mathrm{SR}}^m \gamma_{\mathrm{SR}}^m\right) + \sum_k \sum_n \beta_k^n \log_2\left(1+\frac{p_{\mathrm{RD}}^n \gamma_{\mathrm{RD}}^n}{\beta_k^n}\right)} \tag{C.3}$$

subject to (5.12) and (5.13). Then, the problem can be converted into

$$\underset{p_{\mathrm{SR}}^m, p_{\mathrm{RD}}^n}{\operatorname{argmin}} \frac{1}{2}\left(\frac{1}{\sum_m \log_2\left(1+p_{\mathrm{SR}}^m \gamma_{\mathrm{SR}}^m\right)} + \frac{1}{\sum_k \sum_n \beta_k^n \log_2\left(1+\frac{p_{\mathrm{RD}}^n \gamma_{\mathrm{RD}}^n}{\beta_k^n}\right)}\right) \tag{C.4}$$

subject to (5.12) and (5.13).

Consider functions $u_1(p_{\mathrm{SR}}^1, ..., p_{\mathrm{SR}}^N)$ and $u_2(\beta_1^1, ..., \beta_K^N, p_{\mathrm{RD}_1}^1, ..., p_{\mathrm{RD}_K}^N)$, which are

defined as

$$u_1(p_{\mathrm{SR}}^1, ..., p_{\mathrm{SR}}^N) = \sum_m \log_2 \left(1 + p_{\mathrm{SR}}^m \gamma_{\mathrm{SR}}^m\right) \tag{C.5}$$

and

$$u_2(\beta_k^n, p_{\mathrm{RD}_k}^n) = \beta_k^n \log_2 \left(1 + \frac{p_{\mathrm{RD}}^n \gamma_{\mathrm{RD}_k}^n}{\beta_k^n}\right) \tag{C.6}$$

The Hessian matrix [79] of (C.5) is given by

$$\mathbf{H}\left(u_1\left(p_{\mathrm{SR}}^1, ..., p_{\mathrm{SR}}^N\right)\right) = diag\left(-\left(\frac{\gamma_{\mathrm{SR}}^1}{1 + p_{\mathrm{SR}}^1 \gamma_{\mathrm{SR}}^1}\right)^2, \right.$$
$$\left. \cdots, -\left(\frac{\gamma_{\mathrm{SR}}^N}{1 + p_{\mathrm{SR}}^N \gamma_{\mathrm{SR}}^N}\right)^2\right) \tag{C.7}$$

where $diag(\cdots)$ denotes a diagonal matrix whose diagonal elements are listed in the parentheses. $\mathbf{H}\left(u_1\left(p_{\mathrm{SR}}^1, ..., p_{\mathrm{SR}}^N\right)\right)$ is a negative definite matrix, implying that $u_1(p_{\mathrm{SR}}^1, ..., p_{\mathrm{SR}}^N)$ is a concave function.

The Hessian matrix of (C.6) is

$$\mathbf{H}(u_2(\beta_k^n, p_{\mathrm{RD}_k}^n)) = \frac{p_{\mathrm{RD}_k}^n (\gamma_{\mathrm{RD}_k}^n)^2}{(\beta_k^n + p_{\mathrm{RD}_k}^n \gamma_{\mathrm{RD}+k}^n)^2} \begin{pmatrix} -\frac{p_{\mathrm{RD}_k}^n}{\beta_k^n} & 1 \\ 1 & -\frac{\beta_k^n}{p_{\mathrm{RD}_k}^n} \end{pmatrix} \tag{C.8}$$

Since $p_{\mathrm{RD}_k}^n$ is positive, it is straightforward to derive that $\mathbf{H}(u_2(\beta_k^n, p_{\mathrm{RD}_k}^n))$ is a negative semidefinite matrix, indicating that $u_2(\beta_k^n, p_{\mathrm{RD}_k}^n)$ is a concave function. Therefore, $\sum_k \sum_n u_2(\beta_k^n, p_{\mathrm{RD}_k}^n)$ is also a concave function.

Define function $v(x) = 1/x$, whose first and second derivatives are $v'(x) = -1/x^2$ and $v''(x) = 2/x^3$, respectively. It can be derived that $v(x)$ is a decreasing convex function when $x > 0$. Since $u_1(p_{\mathrm{SR}}^1, ..., p_{\mathrm{SR}}^N)$ is positive and concave, $v(u_1(p_{\mathrm{SR}}^1, ..., p_{\mathrm{SR}}^N)) = 1/\sum_m \log_2 (1 + p_{\mathrm{SR}}^m \gamma_{\mathrm{SR}}^m)$ is a convex function [79]. Similarly, $1/\sum_n \log_2 (1 + p_{\mathrm{RD}}^n \gamma_{\mathrm{RD}}^n)$ is also a convex function. Thus, (C.4) is a convex function overall. Therefore, there is one and only one optimal solution to the objective function of ARA.

Bibliography

[1] D. Tse and P. Viswanath, *Fundamentals of Wireless Communication.* Cambridge University Press, 2005.

[2] V. K. N. Lau and Y. K. R. Kwok, *Channel-Adaptive Technologies and Cross-Layer Designs for Wireless Systems with Multiple Antennas: Theory and Applications.* Wiley, 2006.

[3] J. Jang and K. B. Lee, "Transmit power adaptation for multiuser OFDM systems," *IEEE J. Sel. Areas Commun.*, vol. 21, pp. 171–178, Feb. 2003.

[4] X. Gao, G. Wu, and T. Miki, "End-to-end QoS provisioning in mobile heterogeneous networks," *IEEE Trans. Wireless Commun.*, vol. 11, pp. 24–34, Jun. 2004.

[5] C. Y. Wong, R. S. Cheng, K. B. Letaief, and R. D. Murch, "Multiuser OFDM with adaptive subcarrier, bit and power allocation," *IEEE J. Sel. Areas Commun.*, vol. 17, pp. 1747–1758, Oct. 1999.

[6] T. C. Y. Ng and W. Yu, "Joint optimization of relay strategies and resource allocations in cooperative celluar networks," *IEEE J. Sel. Areas Commun.*, vol. 25, pp. 328–339, Feb. 2007.

[7] R. W. Chang, "Synthesis of band-limited orthogonal signals for multichannel data transmission," *Bell Syst. Tech. J.*, vol. 46, pp. 1775–1796, Dec. 1966.

[8] D. S. W. Hui, V. K. N. Lau, and W. H. Lam, "Cross-layer design for OFDMA wireless systems with heterogeneous delay requirements," *IEEE Trans. Wireless Commun.*, vol. 6, pp. 2872–2880, Aug. 2007.

[9] G. C. Song and Y. Li, "Utility-based resource allocation and scheduling in OFDM-based wireless broadband networks," *IEEE Commun. Mag.*, vol. 43, pp. 127–134, Dec. 2005.

[10] J. Chuang and N. Sollenberger, "Beyond 3G:wideband wireless data access based on OFDM and dynamic packet assignment," *IEEE Commun. Mag.*, vol. 38, pp. 78–87, Jul. 2000.

[11] E. C. van der Meulen, "Three-terminal communication channels," *Adv. Appl. Prob.*, vol. 3, pp. 120–154, 1971.

[12] J. N. Laneman, D. N. C. Tse, and G. W. Wornell, "Cooperative diversity in wireless networks: efficient protocols and outage behavior," *IEEE Trans. Inform. Theory*, vol. 50, pp. 3062–3080, Dec. 2004.

[13] T. S. Rappaport, *Wireless Communications: Principles and Practice.* Pearson Education, 2002.

[14] D. N. Knisely, S. Kumar, S. Laha, and S. Nanda, "Evolution of wireless data services: IS-95 to CDMA2000," *IEEE Commun. Mag.*, vol. 36, pp. 140–149, Oct. 1998.

[15] R. Attar, R. Ghosh, C. Lott, M. X. Fan, P. Black, R. Rezaiifar, and P. Agashe, "Evolution of CDMA2000 cellular networks: multicarrier EV-DO," *IEEE Commun. Mag.*, vol. 44, pp. 46–53, Mar. 2006.

[16] Technology of high speed packet access (HSPA). [Online]. Available: http://www.nomor.de/uploads/b0/2m/b02mwrVvIa5ZUtVrFeSP1w/Technology_of_HSPA.pdf

[17] J. Bergman, M. Ericson, D. Gerstenberger, B. Göransson, J. Peisa, and S. Wager, "HSPA evolution — Boosting the performance of mobile broadband access," *Ericsson Review*, vol. 85:1, pp. 32–37, 2008.

[18] 3GPP Technical Report 36.913 Version 9.0.0, Requirements for further advancements for Evolved Universal Terrestrial Radio Access (E-UTRA) (LTE-Advanced) (Release 9), Dec. 2009.

[19] K. Doppler, S. Redana, M.Wódczak, P. Rost, and R. Wichman, "Dynamic resource assignment and cooperative relaying in cellular networks: Concept and performance assessment," *EURASIP Journal on Wireless Commun. and Networking*, vol. 2009, pp. 1–14, 2009.

[20] R. U. Nabar and H. B. F. W. Kneubuhler, "Fading relay channels: performance limits and space-time signal design," *IEEE J. Sel. Areas Commun.*, vol. 22, pp. 1099–1109, Aug. 2004.

[21] A. J. Goldsmith and S. G. Chua, "Variable-rate variable-power MQAM for fading channel," *IEEE Trans. Commun.*, vol. 45, pp. 1218–1230, Oct. 1997.

[22] T. J. Willink and P. H. Wittke, "Optimization and performance evaluation of multicarrier transmission," *IEEE Trans. Inform. Theory*, vol. 43, pp. 426–440, Aug. 1997.

[23] G. C. Song, "Cross-layer resource allocation and scheduling in wireless multicarrier networks," Ph.D. dissertation, Georgia Institute of Technology, 2005.

[24] R. Knopp and P. A. Humblet, "Information capacity and power control in single-cell multiuser communications," in *Proc. IEEE ICC'05*, Seattle, U.S.A., Jun. 1995, pp. 331–335.

[25] G. Caire, G. Taricco, J. Ventura-Traveset, and E. Biglieri, "A multiuser approach to narrowband cellular communications," *IEEE Trans. Inform. Theory*, vol. 43, pp. 1503–1517, 1997.

[26] M. Stojanovic and Z. Zvonar, "Performance of multiuser diversity reception in rayleigh fading CDMA channels," *IEEE Trans. Commun.*, vol. 47, pp. 356–359, Mar. 1999.

[27] M. Grossglauser and D. Tse, "Mobility increases the capacity of ad hoc wireless networks," *IEEE/ACM Trans. Networking*, vol. 10, pp. 477–486, Aug. 2002.

[28] A. Sang, X. D. Wang, M. Madihian, and R. D. Gitlin, "A flexible downlink scheduling scheme in cellular packet data systems," *IEEE Trans. Wireless Commun.*, vol. 5, pp. 568–577, Mar. 2006.

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

[30] 3GPP Technical Specification 25.308 Version 5.2.0, High Speed Downlink Packet Access (HSDPA): Overall description, Mar. 2002.

[31] 3GPP Technical Report 25.848 Version 4.0.0, Physical layer aspects of UTRA High Speed Downlink Packet Access, Mar. 2001.

[32] K. P. Ho and J. M. Kahn, "Transmission of analog signals using multicarrier modulation: a combined source-channel coding approach," *IEEE Trans. Commun.*, vol. 44, pp. 1432–1443, Nov. 1996.

[33] T. J. Willink and P. H. Wittke, "Optimization and performance evaluation of multicarrier transmission," *IEEE Trans. Inform. Theory*, vol. 43, pp. 426–440, Mar. 1997.

[34] I. Kalet, "The multitone channel," *IEEE Trans. Commun.*, vol. 37, pp. 119–124, Feb. 1989.

[35] P. W. C. Chan, E. S. Lo, R. R. Wang, E. K. S. Au, V. K. N. Lau, R. S. Cheng, W. H. Mow, R. D. Murch, and K. B. Letaief, "The evolution path of 4G networks: FDD or TDD?" *IEEE Commun. Mag.*, vol. 44, pp. 42–50, Dec. 2006.

[36] W. J. Rhee and J. M. Cioffi, "Increase in capacity of multiuser OFDM system using dynamic subchannel allocation," in *Proc. IEEE VTC Spring'00*, Tokyo, Japan, 2000.

[37] D. Kivanc, G. Q. Li, and H. Liu, "Computationally efficient bandwidth allocation and power control for OFDMA," *IEEE Trans. Wireless Commun.*, vol. 2, pp. 1150–1158, Nov. 2003.

[38] Z. K. Shen and J. G. Andrews, "Adaptive resource allocation in multiuser OFDM systems with proportional rate constraints," *IEEE Trans. Wireless Commun.*, vol. 4, pp. 2726–2737, Nov. 2005.

[39] T. M. Cover and J. A. Thomas, *Elements of Information Theory.* Wiley, 1991.

[40] N. Zhou, X. Zhu, Y. Huang, and H. Lin, "Novel batch dependant cross-layer scheduling for multiuser OFDM systems," in *Proc. IEEE ICC'08*, Beijing, China, May 2008, pp. 3878–3882.

[41] M. Herdin, "A chunk based OFDM amplify-and-forward relaying scheme for 4G mobile radio systems," in *Proc. IEEE ICC'06*, vol. 10, Istanbul, Turkey, Jun. 2006, pp. 4507–4512.

[42] I. Hammerstroem and A. Wittneben, "On the optimal power allocation for nonregenerative OFDM relay links," in *Proc. IEEE ICC'06*, vol. 10, Istanbul, Turkey, Jun. 2006, pp. 4463–4468.

[43] W. Y. Wang, S. F. Yan, and S. Y. Yang, "Optimally joint subcarrier matching and power allocation in OFDM multihop system," *EURASIP Journal on Advances in Signal Processing*, vol. 8, Jan. 2008.

[44] Y. Wang, X. C. Qu, T. Wu, and B. L. Liu, "Power allocation and subcarrier pairing algorithm for regenerative OFDM relay system," in *Proc. IEEE VTC Spring 07*, Dublin, Ireland, Apr. 2007, pp. 2727–2731.

[45] G. Q. Liu and H. Liu, "On the capacity of broadband relay networks," in *Proc. IEEE ACSSC'04*, vol. 2, Pacific Grove, USA, Nov. 2004, pp. 1318–1322.

[46] N. Zhou, X. Zhu, Y. Huang, and H. Lin, "Adaptive resource allocation for multi-destination relay systems based on OFDM modulation," in *Proc. IEEE ICC'09*, Dresden, Germany, Jun. 2009, pp. 1–5.

[47] Y. W. Pan, A. Nix, and M. Beach, "A game theoretic approach to distributed resource allocation for OFDMA-based relaying networks," in *Proc. IEEE PIMRC'08*, Cannes, France, Sep. 2008, pp. 1–5.

[48] Y. Sun, Y. Z. Xiao, M. Zhao, X. F. Zhong, S. D. Zhou, and N. B. Shroff, "Joint power and channel resource allocation for F/TDMA decode and forward relay networks," in *Proc. IEEE GLOBECOM'09*, Hawaii, U.S.A., Nov. 2009, pp. 1–6.

[49] W. Nam, W. Chang, S. Y. Chung, and Y. H. Lee, "Transmit optimization for relay-based cellular OFDMA systems," in *Proc. IEEE ICC'07*, Glasgow, Scotland, Jun. 2007, pp. 5714–5719.

[50] N. Zhou, X. Zhu, Y. Huang, and H. Lin, "Optimal asymmetric resource allocation with limited feedback for OFDM based relay systems," *IEEE Trans. Wireless Commun.*, vol. 9, pp. 552–557, Feb. 2010.

[51] D. Veronesi, S. Tomasin, and N. Benvenuto, "Cross-layer optimization for multimedia traffic in CDMA cellular networks," *IEEE Trans. Wireless Commun.*, vol. 7, pp. 1379–1388, Apr. 2008.

[52] E. K. Bowdon, "Priority assignment in a network of computers," *IEEE Trans. Computers*, vol. C-18, pp. 1021–1026, Nov. 1969.

[53] F. X. Zhang and A. Burns, "Schedulability analysis for real-time systems with EDF scheduling," *IEEE Trans. Computers*, vol. 58, pp. 1250–1258, Sep. 2009.

[54] B. Maglaris and M. Schwartz, "Performance evaluation of a variable frame multiplexer for integrated switched networks," *IEEE Trans. Commun.*, vol. 29, pp. 800–807, Jun. 1981.

[55] R. R. Pillai, "A distributed overload control algorithm for delay-bounded call setup," *IEEE/ACM Trans. Networking*, vol. 9, pp. 780–789, Dec. 2001.

[56] J. Nagle, "On packet switches with infinite storage," *IEEE Trans. Commun.*, vol. 35, pp. 435–438, Apr. 1987.

[57] A. K. Parekh and R. G. Gallager, "A generalized processor sharing approach to flow control in integrated services networks: the single-node case," *IEEE/ACM Trans. Networking*, vol. 1, pp. 344–357, Jun. 1993.

[58] ——, "A generalized processor sharing approach to flow control in integrated services networks: the multiple node case," *IEEE/ACM Trans. Networking*, vol. 2, pp. 137–150, Apr. 1994.

[59] A. Demers, S. Keshav, and S. Shenker, "Analysis and simulation of a fair queueing algorithm," in *Proc. of ACM SIGCOMM'89*, Austin, U.S.A., Sep. 1989, pp. 1–12.

[60] A. Martinez, G. Apostolopoulos, F. J. Alfaro, J. L. Sanchez, and J. Duato, "Efficient deadline-based qos algorithms for high-performance networks," *IEEE Trans. Computers*, vol. 57, pp. 928–939, Jul. 2008.

[61] S. Ng and J. Mark, "Multiaccess model for packet switching with a satellite having processing capability: Delay analysis," *IEEE Trans. Commun.*, vol. 26, pp. 283–290, Feb. 1978.

[62] E. L. Hahne, "Round-robin scheduling for max-min fairness in data networks," *IEEE J. Sel. Areas Commun.*, vol. 9, pp. 1024–1039, Sep. 1991.

[63] M. Shreedhar and G. Varghese, "Efficient fair queueing using deficit round-robin," *IEEE/ACM Trans. Networking*, vol. 4, pp. 375–385, Jun. 1996.

[64] C. L. Liu and J. W. Layland, "Scheduling algorithm for multiprogramming in a hard real-time environment," *J. ACM*, vol. 20, pp. 40–61, Jan. 1973.

[65] L. Georgiadis, R. Guerin, and A. Parekh, "Optimal multiplexing on a single link: Delay and buffer requirements," in *Proc. IEEE INFOCOM'94*, Jun. 1994, pp. 524–532.

[66] K. B. Letaief and Y. J. Zhang, "Dynamic multiuser resource allocation and adaptation for wireless systems," *IEEE Trans. Wireless Commun.*, vol. 13, pp. 38–47, Aug. 2006.

[67] C. Verikoukis, L. Alonso, and T. Giamalis, "Cross-layer optimization for wireless systems: A european research key challenge," *IEEE Commun. Mag.*, vol. 43, pp. 1–3, Jul. 2005.

[68] G. Song and Y. Li, "Cross-layer optimization for OFDM wireless networks-part I: theoretical framework," *IEEE Trans. Wireless Commun.*, vol. 4, pp. 614–624, Mar. 2005.

[69] ——, "Cross-layer optimization for OFDM wireless networks-part II: algorithm development," *IEEE Trans. Wireless Commun.*, vol. 4, pp. 625–634, Mar. 2005.

[70] H. Zuleita, V. Lau, and R. S. K. Cheng, "Cross-layer design of FDD-OFDM systems based on ACK/NAK feedbacks," *IEEE Trans. Inform. Theory*, vol. 55, pp. 4568–4584, Oct. 2009.

[71] H. J. Kushner and P. A. Whiting, "Convergence of proportional-fair sharing algorithms under general conditions," *IEEE Trans. Wireless Commun.*, vol. 3, pp. 1250–1259, Jul. 2004.

[72] M. Andrews, K. Kumaran, K. Ramanan, A. Stolyar, P. Whiting, and R. Vijayakumar, "Providing quality of service over a shared wireless link," *IEEE Commun. Mag.*, vol. 2, pp. 150–154, Feb. 2001.

[73] Q. W. Liu, X. Wang, and G. B. Giannakis, "A cross-layer scheduling algorithm with QoS support in wireless networks," *IEEE Trans. Veh. Technol.*, vol. 55, pp. 839–847, May 2006.

[74] S. Shakkottai and A. L. Stolyar, "Scheduling for multiple flows sharing a time-varying channel: the exponential rule," *AMC Translations*, vol. 207, pp. 185–202, 2002.

[75] F. P. Kelly, "Charging and rate control for elastic traffic," *Eur. Trans. Telecommun.*, vol. 8, pp. 33–37, 1997.

[76] Y. Yu and W. Zhou, "Resource allocation for OFDMA system based on genetic algorithm," in *Proc. IEEE IWCLD'07*, vol. 10, Tai'an, China, Sep. 2007, pp. 65–69.

[77] K. B. Johnsson and D. C. Cox, "An adaptive cross-layer scheduler for improved QoS support of multiclass data services on wireless systems," *IEEE J. Sel. Areas Commun.*, vol. 23, pp. 334–343, Feb. 2005.

[78] J. Bentley, *Programming Pearls.* Addison-Wesley, 1999.

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

[80] J. D. C. Little, "A proof of the queueing formula $l = \lambda w$," *Operations Research*, vol. 9, pp. pp. 383–387, May 1961.

[81] L. Georgiadis, M. J. Neely, and L. Tassiulas, *Resource Allocation and Cross-Layer Control in Wireless Networks.* Now Publishers Inc, 2006.

[82] A. Erylimaz, R. Srikant, and J. R. Perkins, "Stable scheduling policies for fading wireless channels," *IEEE/ACM Trans. Network.*, vol. 13, pp. 411–424, Apr. 2005.

[83] T. H. Corman, C. E. Leiserson, R. L. Rivest, and C. Stein, *Introduction to Algorithms.* MIT Press, 2001.

[84] W. Chen, L. Dai, K. B. Letaief, and Z. G. Cao, "A unified cross-layer framework for resource allocation in cooperative networks," *IEEE Trans. Wireless Commun.*, vol. 7, pp. 3000–3012, Aug. 2008.

[85] H. Wu, Y. F. Wang, C. Xiong, and D. C. Yang, "A novel relay selection scheme with simplified power allocation for wireless relay networks," in *Proc. of IEEE GLOBECOM'09*, Honolulu, U.S.A., Nov. 2009, pp. 1–5.

[86] L. Vandendorpe, R. T. Duran, J. Louveaux, and A. Zaidi, "Power allocation for OFDM transmission with DF relaying," in *Proc. IEEE ICC'08*, Beijing, China, May 2008, pp. 3795 – 3800.

[87] M. Ibrahimi and B. Liang, "Efficient power allocation in cooperative OFDM system with channel variation," in *Proc. IEEE ICC'08*, Beijing, China, May 2008, pp. 3022–3028.

[88] Multi-hop relay system evaluation methodology. [Online]. Available: http://ieee802.org/16/relay/docs/80216j-06_013r3.pdf

[89] X. Zhu and R. D. Murch, "Performance analysis of maximum likelihood detection in a MIMO antenna system," *IEEE Trans. Commun.*, vol. 50, pp. 187–191, Feb. 2002.

[90] W. Rudin, *Principles of Mathematical Analysis.* McGraw-Hill, 1964.