5-2016

Quality of Experience Aware Spectrum Efficiency and Energy Efficiency Over Wireless Heterogeneous Networks

Yiran Xu
Utah State University

QUALITY OF EXPERIENCE AWARE SPECTRUM EFFICIENCY AND ENERGY EFFICIENCY OVER WIRELESS HETEROGENEOUS NETWORKS

by

Yiran Xu

A dissertation submitted in partial fulfillment
of the requirements for the degree

of

DOCTOR OF PHILOSOPHY

in

Electrical Engineering

Approved:

Dr. Rose Q. Hu Major Professor	Dr. Todd Moon Committee Member
Dr. Jacob Gunther Committee Member	Dr. Chris Winstead Committee Member
Dr. Xiaojun Qi Committee Member	Dr. Mark R. McLellan Vice President for Research and Dean of the School of Graduate Studies

UTAH STATE UNIVERSITY
Logan, Utah

2016

Abstract

Quality of Experience Aware Spectrum Efficiency and Energy Efficiency over Wireless Heterogeneous Networks

by

Yiran Xu, Doctor of Philosophy

Utah State University, 2016

Major Professor: Dr. Rose Q. Hu
Department: Electrical and Computer Engineering

Propelled by the explosive increases in mobile data traffic volume, existing wireless technologies are stretched to their capacity limits. There is a tremendous need for an expansion in system capacity and an improvement on energy efficiency. In addition, wireless network will support more and more multimedia services and applications, in which user experience has been always an important factor in evaluating the overall network performance. In order to keep pace with this explosion of data traffic and to meet the emerging quality of experience needs, wireless heterogeneous networks have been introduced as a promising network architecture evolution of the traditional cellular network.

In this dissertation, we explore video quality-aware spectrum efficiency and energy efficiency in wireless heterogeneous networks—the potentials and the associated technical challenges. In particular, aiming to significantly enhance spectrum efficiency, we need to tackle the interference issue, which is exacerbated in heterogeneous network due to ultra dense node deployment as well as heterogeneity nature of various nodes. Specifically, we first study an optimal intra-cell inter-tier cooperation to mitigate interference between high power nodes and low power nodes. Together with cooperation, optimal mobile association and resource allocation schemes are also intensively investigated in heterogeneous network to

achieve system load balancing so that bandwidth at high power and low power nodes can be utilized in the optimal way. The proposed scheme can greatly alleviate inter-tier interference and significantly increase overall system spectrum efficiency in a heterogeneous network. We then further apply advanced algorithms such as precoding, and non-orthogonal multiple access into intra-cell inter-tier cooperation so that the overall system spectrum efficiency and user experience are even more improved. When supporting a video type application in such a heterogeneous network, considering only spectrum efficiency is far from enough as video application is bandwidth consuming, battery consuming, and quality demanding. We develop a video quality-aware spectrum and energy efficient resource allocation scheme in a wireless heterogeneous network and propose novel performance metrics to establish fundamental relationships among spectrum efficiency, energy efficiency, and quality of experience. Extensive simulations are conducted to evaluate the trade-off performance among three performance metrics.

(149 pages)

Public Abstract

Quality of Experience Aware Spectrum Efficiency and Energy Efficiency over Wireless Heterogeneous Networks

by

Yiran Xu, Doctor of Philosophy

Utah State University, 2016

Major Professor: Dr. Rose Q. Hu
Department: Electrical and Computer Engineering

At the turn of the 21st century, people experienced a revolution in consumer electronics and telecommunication technologies. The smart phone changed the Internet landscape in a way no other technology has in the last decade. The widespread popularity of multimedia-friendly connected devices like smart phones and tablets is triggering explosive mobile application proliferation and data traffic growth. Service providers are struggling to keep pace with the rapidly increasing demands from customers. Legions of consumers are embracing these innovative devices, and their hunger for more and more bandwidth and quality of experience is eating up peak-time bandwidth and heaping pressure on current cellular networks. Based on the forecast data, global mobile traffic grew 69% in 2014, which was nearly 30 times the size of the entire global Internet in 2000, and it will increase nearly 10-fold by 2019. In contrast, the average data speed will only increase 19% annually in the next five years. Clearly there exists a huge gap between the growth rate from air interface technologies and the growth rate of customer needs. To maintain mobile service profitability, and narrow the gap between increasing demands and scarce network resources, it is necessary to explore the potential benefits of novel network architecture and cutting

edge wireless technologies simultaneously. There are two major tendencies in this cellular revolution: cellular network topology shift and evolution of wireless technologies.

One of the interesting trends is to shift cellular topology and architecture by introducing heterogeneity. In heterogeneous networks, small cells are deployed along with macro-cells to expand coverage range and improve spatial reuse. Specifically, the base station located in a small cell has a relatively lower transmit power but has the same spectrum capacity as the base station in a macro-cell. The higher the deployment density, the better chance that user equipment can be served by a nearby base station with strong signal strength. Thereby, with the deployment of inexpensive low power base stations through the use of small cells, network capacity, spectrum efficiency, and energy efficiency can be improved considerably.

The other tendency in this cellular revolution is to explore new features of novel wireless technologies and standards. A number of researchers have investigated new radio access techniques, radio resource allocation, cooperative transmission schemes, and so on. All of these innovative ideas aim to mitigate the interfering signals and enhance the desired signal strength to create good quality of service for the end users.

In this dissertation, we will provide an overview of wireless heterogeneous networks and current state-of-the-art wireless technologies. In particular, we explore radio resource allocation, cooperative transmission, precoder design, and multiple access schemes in downlink heterogeneous networks, and study their impacts on system performance and user experience. Furthermore, we take video applications into account and investigate the potential of heterogeneous networks in video quality-aware transmission.

To my parents for their love and support.

Acknowledgments

First, I would like to thank my supervisor, Professor Rose Qingyang Hu, for her invaluable support, inspiration, and instruction throughout my study at Utah State University. I have learned tremendously from her insightful comments and constructive criticism, which greatly enhanced my research and will continue to benefit my future career development. I would also like to thank Professors Todd Moon, Jacob Gunther, Chris Winstead, Xiaojun Qi, YangQuan Chen, and Anthony Chen for teaching me in class and serving as my committee members. During my Ph.D. study, Prof. Gunther helped me tackle mathematical problems either in class or in my own research. The discussions with him were always helpful and insightful.

Next, I would like to thank Professor Yi Qian, Professor Taieb Znati, Dr. Geng Wu, Dr. Clara Qian Li, and Dr. Lili Wei, who worked closely with me on many research projects. Their collaboration and discussions inspired my new ideas and helped me solve the technical problems in my research efficiently.

I am also very grateful to my colleagues and friends at Utah State University. In addition to those listed above, I would like to thank Professor Xianfu Lei, Dr. Bei Xie, Dr. Tao He, Dr. Junlin Zhang, Xue Chen, Zhengfei Rui, Zekun Zhang, Haijian Sun, Xuan Xie, David Neal, Dr. Zhouyuan Li, and Zhuo Li. They made my life in Logan much more enjoyable.

In addition, I would like to thank Prof. I-Tai Lu, Dr. Enoch Lu, Dr. Jiang Chang, Dr. Xiao Han, Dr. Jialing Li, Dr. Sha Hua, Dr. Zhan Ma, and Fanyi Duanmu for their help with my research.

I would also like to thank Dr. Ming Zhang, Braden Gibson, Ryan Zenker, and David Scherer for offering me a chance to intern at EMC Corporation and providing selfless help during my internship.

Finally, I particularly want to thank my parents for their endless love and support, and my wife Bingyi Xiang for her persistent encouragement and company. Without their love and encouragement, it would be impossible for me to gain such achievements.

This work is supported by National Science Foundation.

Yiran Xu

Contents

List of Tables

List of Figures

Notation

N_c	number of macro-cells
N_r	number of pico-cells or relay nodes per macro-cell
N_p	total number of number of pico-cells or relay nodes
N_u	number of UEs
P_m	transmit power of mBS
P_p	transmit power of pBS
P_r	transmit power of relay node
P_t	transmit power
P_c	circuit power
P_s	static part in circuit power
γ	signal-to-interference-noise ratio
h	channel gain
x	mobile association variable
n	resource allocation variable
R	received data rate
$\mathcal{L}$	Lagrange function
ζ	drain efficiency of power amplifier
ξ	scale factor of bandwidth-dependent function
C	total bandwidth resource
$\mathbf{A} \circ \mathbf{B}$	Hadamard product of matrix $\mathbf{A}$ and $\mathbf{B}$
$Diag\{\mathbf{A}\}$	diagonal column vector of matrix $\mathbf{A}$
$Tr\{\mathbf{A}\}$	trace of matrix $\mathbf{A}$
$\mathbf{A}^*$	conjugate transpose of matrix $\mathbf{A}$
$\mathbf{A}^{-1}$	inverse of matrix $\mathbf{A}$
$\mathbb{E}\{f(x)\}$	ensemble average of function $f(x)$ over the pdf of the random variable x.

Acronyms

BS	Base Station
RN	Relay Node
mBS	macro-BS
pBS	pico-BS
UE	User Equipment
MUE	Macro-UE
PUE	Pico-UE
RUE	Relay-UE
CUE	Cooperative UE
RB	Resource Block
3GPP	3rd Generation Partnership Project
LTE-A	Long-Term Evolution Advanced
OFDM	Orthogonal Frequency-Division Multiplexing
OFDMA	Orthogonal Frequency-Division Multiple Access
PtP	Point-to-Point
CoMP	Coordinated Multipoint Processing
CSI	Channel State Information
RAN	Radio Access Network
D2D	Device-to-Device
THP	Tomlinson-Harashima precoding
NOMA	Non-Orthogonal Multiple Access
SIC	Successive Interference Cancellation
CT	Cooperative Transmission
MU-MIMO	Multi-User Multipl-Input and Multiple-Output
DFT	Discrete Fourier Transform
DPC	Dirty Paper Coding
GA	Genetic Algorithm
SE	Spectrum/Spectral Efficiency
EE	Energy Efficiency
QoS	Quality of Service
QoE	Quality of Experience
QSE	Quality-based SE
QEE	Quality-based EE
MOOP	Multi-Objective Optimization Problem
SINR	Signal-to-Interference-Noise Ratio
PSNR	Peak Signal-to-Noise Ratio
AWGN	Additive White Gaussian Noise
KKT conditions	Karush-Kuhn-Tucker conditions
MINLP	Mixed Integer Nonlinear Programming problem

Chapter 1

Introduction

1.1 Challenges and Motivations

The widespread popularity of multimedia-friendly connected devices like smart phones and tablets is triggering explosive mobile video consumption and data traffic growth. Service providers are struggling to keep pace with the rapidly increasing demands from customers. Legions of consumers are embracing these innovative devices, and their hunger for multimedia content delivery and quality of experience (QoE) is eating up peak-time bandwidth and heaping pressure on current cellular networks. To maintain mobile service profitability, and narrow the gap between increasing demands and scarce network resources, it is necessary to explore the potential benefits of novel network architecture and cutting edge wireless technologies simultaneously.

In traditional cellular networks, a base station (BS) consumes a significant amount of energy to support the activities of user equipment (UE), especially cell edge users. Emerging high-density, heterogeneous wireless networks introduce a hierarchical infrastructure, where high power BSs provide blanket coverage and seamless mobility, while low power nodes, such as femto- and pico-BS, help support cell edge users and boost cell capacity [1–4]. Usually deployed at coverage holes or capacity-demanding hotspots, these low power nodes can extend the wireless service coverage range and expand the cell capacity.

In this dissertation, we will investigate QoE-aware spectrum efficient and energy efficient mobile association and resource allocation schemes in wireless heterogeneous networks. The main objective is to explore advanced mobile association and resource allocation in wireless heterogeneous networks, focusing on the interplay among spectrum efficiency (SE), energy efficiency (EE), and QoE. We first study some cutting edge wireless technologies, such as Coordinated Multipoint Processing (CoMP), Tomlinson-Harashima precoding (THP),

and Dirty Paper Coding (DPC), and their applications in heterogeneous networks. Then we further extend our work to a video delivery network with QoE requirements. We propose two new performance metrics by taking video quality into account and construct the trade-off relationships among bandwidth consumption, power consumption and perceived video qualities. These metrics allow us to obtain profound insights on system-wide spectrum efficiency and energy efficiency from the perspective of video quality.

1.2 Wireless Heterogeneous Networks

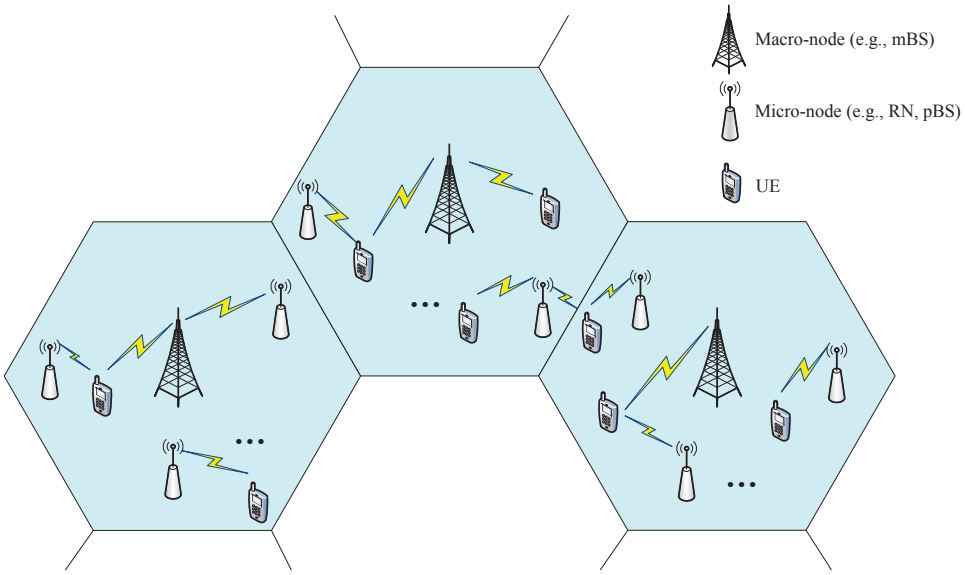

Fig. 1.1: Wireless heterogeneous network

As a promising paradigm in next generation networks, wireless heterogeneous networks bring heterogeneity into the network architecture. Specifically, we consider a two-tier downlink wireless heterogeneous network as shown in Fig. 1.1. Each cell is divided into several sectors, where one macro-node, e.g., mBS, and multiple micro-nodes, e.g., relay node (RN) and pBS, are deployed in each cell simultaneously. To differentiate from the macro-cells that are created by macro-nodes, cells created by the micro-nodes are called micro-cells. Compared to a macro-node, a micro-node typically transmits at a low power level and acts like a fully-featured mini-BS. Their reduced size and cost make them easily deployed for improving conditions in coverage holes and providing higher data rates at cell edge or in

hotspots. UEs are uniformly distributed in the network, so that each UE can be served by either a macro-node or a micro-node, depending on the location and service requirement of the UE. The deployment of micro-nodes can contribute to the following benefits:

- **Expanded network coverage** : The deployment of micro-nodes introduces smaller cells on top of the conventional cellular system and effectively expands the cellular network coverage.
- **Increased network capacity** : Micro-nodes act like a fully-featured mini-BS. The deployment of micro-nodes increases network density so that it can serve more UEs, resulting in an increase of network capacity.
- **Enhanced user performance** : By deploying micro-nodes, the distance between BS and UE is shortened, giving the UE a stronger signal from the BS, resulting in a higher data transfer rate and better performance.
- **Improved energy efficiency** : Micro-nodes have relatively lower transmit power. Thus, when deploying micro-notes, it is not necessary to increase the macro-node's transmit power to serve cell edge users, resulting in less power consumption and greater energy efficiency.
- **Lower the cost**: Micro-nodes are relatively inexpensive. By deploying micro-nodes instead of increasing the number of expensive macro-nodes, people can lower the networks operational expenditures.

Challenges always come with opportunities. Aside from the aforementioned benefits, heterogeneous deployment also causes some technical challenges during implementation. In this dissertation, we mainly focus on the following three challenges:

- **Inter-cell and intra-cell interference**: The interference coordination problem is significantly more challenging in a wireless heterogeneous network. In addition to inter-cell interference, cells from different layers, i.e., macro- or micro-layers, have

different transmit powers and are overlaid on each other, resulting in new and complicated interference scenarios.

- **Load balancing**: Due to the disparity between the transmit power of the macro-node and that of the micro-node, if a micro-node is not placed specifically in a hot spot, only a small number of UEs will connect to the micro-node, which will limit the gain from offloading the traffic from the macro-cells.

- **Mobile association and resource allocation**: Traditionally, the best power association scheme sacrifices load balancing for interference mitigation, while a range expansion scheme can achieve load balancing but creates strong interference for cell edge users. Also, in a large-scale system, user fairness is thought of as an important metric. Therefore, the goal of joint mobile association and resource allocation schemes is to maximize system performance, achieve the tradeoff between load balancing and interference, and also guarantee user fairness.

1.3 Dissertation Outline

In this dissertation, we will focus on dealing with the aforementioned challenges in heterogeneous networks and showing the performance improvements. In particular, we will analyze the problems explicitly and propose effective schemes to tackle them. By conducting system-level simulations, we will evaluate our proposed schemes in terms of various performance metrics.

The dissertation is organized as follows.

In Chapter 2, we introduce a cooperative transmission to combat the intra-cell and inter-cell interferences in a relay-based heterogeneous network. Specifically, we formulate an optimization problem to maximize the log-scale throughput function and investigate optimal mobile association and resource allocation strategies to improve cell edge performance and ensure users' proportional fairness.

In Chapter 3, we extend our work in the previous chapter by combining a precoding technique with CoMP to further increase data transfer rates for end users. In the proposed resource allocation framework, we first employ Tomlinson-Harashima precoding to cancel out inter-user interferences so that mBS and pBS can serve multiple cell edge UEs simultaneously, resulting in a more efficient systemic utilization of radio resources.

We then propose, in Chapter 4, a hybrid multi-user multiple-input and multiple-output (MU-MIMO) and non-orthogonal multiple access (NOMA) design scheme in wireless heterogeneous networks to improve the system throughput and also to increase multi-user diversity gains by exploiting the heterogeneous nature of the supporting wireless networks. The best user cluster is formed in a NOMA group and then a precoding based MU-MIMO scheme is applied to NOMA composite signals. The problem is further formulated as a resource scheduling optimization problem with proportional fairness purpose. Aiming to ensure the global optimality, a brute-force search algorithm is used to solve the problem.

In Chapter 5, we further explore NOMA scheme with successive interference cancellation (SIC) in a multi-antenna system. The formulated system model can be regarded as two MU-MIMO sub-systems, and NOMA-SIC is applied on the receiving side. Aiming to improve the system capacity and increase data transmission, we propose a cooperative NOMA scheme and formulate a joint mobile association and resource scheduling optimization problem. Genetic algorithm is implemented to solve the problem efficiently.

We consider video applications over heterogeneous networks in Chapter 6. In particular, we focus on the interplay between video quality and resource consumption. To this end, we propose two new system performance metrics, video quality-aware spectral efficiency (QSE) and video quality-aware energy efficiency (QEE), which measure the video quality per unit of radio resource consumption and per unit of power consumption, respectively. Based on the new performance metrics, a joint optimization problem is formulated to derive mobile association and resource allocation schemes for video connections in a wireless heterogeneous network. Furthermore, we formulate a multi-objective optimization problem (MOOP) to investigate the tradeoffs and interplay between QSE and QEE in the heterogeneous network.

In Chapter 7, we propose a multi-objective optimization framework to address the joint mobile association and resource allocation problem in a video transmitted wireless heterogeneous network. We consider user QoE as one of the design objectives together with two other performance metrics to characterize the design tradeoffs among perceived video quality, power consumption, and network resource consumption.

Chapter 8 concludes the thesis and discusses some directions for future research.

To help the understanding, we first summarize the notations and abbreviations frequently used throughout this dissertation in **xvi** and **xvii**.

Chapter 2

Optimal Intra-cell Cooperation in Heterogeneous Relay Networks

2.1 Introduction

Driven by the proliferation of wireless devices and applications, future wireless systems are required to support various applications at a much higher capacity and a higher spectral efficiency. Based on the forecast data, global mobile traffic increases 66x with an annual growth rate of 131% between 2008 and 2013. In contrast, the peak data rate from 3G UMTS to 4G LTE-A only increases 55% annually [5]. Clearly there exists a huge gap between the growth rate of new air interface and the growth rate of customer needs. In order to narrow such a gap fundamentally, it is necessary to make changes from infrastructure aspect, as today's wireless link efficiency is approaching its Shannon limit. Therefore, heterogeneous network with BS of diverse sizes and capabilities has been considered as a mainstream technology for the future wireless network.

Recently, cooperative transmission (CT), a promising technology used in 3GPP LTE-A, has been extensively investigated to further improve the cell edge performance [6]. [7–9] explored CT in the traditional homogeneous networks. Simulation results revealed that CT can tremendously improve the homogeneous system performance. [10–12] mainly investigated the resource allocation solutions for relay-based OFDMA cellular networks. They proposed coordinated resource allocation schemes and showed these schemes can significantly improve the network performance in terms of power saving, user utilities and system throughput. The deployment of heterogeneous networks has created a number of new cell edge scenarios [13], which make CT technology even more attractive in the heterogeneous networks. When implementing relay nodes (RN) in a heterogeneous network, user data is

transmitted via multi-hops on the air interface, i.e. from the donor BS to the RN first and then from RN to the user. So user data is available at both the donor BS and RN and it is possible to implement intra-cell CT at the donor BS and RN within the cell. Such a combination of relay communication in a heterogeneous network and CT has been proposed for consideration in LTE release-11 and beyond [14]. In this chapter, we focus on the intra-cell CT in a heterogeneous network with relays and propose an optimal cooperation scheme that aims to maximize the long-term system throughput as well ensure user fairness.

2.2 Cooperative Transmission in Heterogeneous Networks

We consider downlink communication in a heterogeneous relay network and investigate intra-cell cooperation in such a network. Each cell is divided into several sectors. Each sector has one BS and multiple RNs are deployed in each sector to further increase the capacity and coverage. The BS in each sector is called the donor BS for RNs in the same sector. Communications between a node and a UE can be achieved in three different ways:

(1) direct transmission between BS and UE;

(2) two-hop transmission with RN's help;

(3) cooperative transmission from BS and RN in the same cell to the UE.

Usually cooperative transmission involves extensive data exchange and high signaling overhead between different nodes. However, in a RN network, since the data packets transmitted from RN to UEs are always available at the donor BS, the CT between the donor BS and RN is made easier. We denote such a cooperation as intra-cell cooperation. In this paper, we want to focus on the optimal design of the intra-cell cooperation, which aims to maximize the long-term log-scale system throughput.

In our system model, we denote the total number of UEs as N_u. They are uniformly distributed in N_c sectors. There are N_r RNs in each sector and the total number of RNs in the network is $N_p = N_r N_c$. The relays studied in this chapter use out of band backhaul. We assume all the BSs have the same transmit power P_m and all the RNs have the same transmit

power P_r, $P_r < P_m$. A RN's footprint is much smaller than that of the donor BS due to RN's lower transmit power. As a result, conventional best-power based mobile association scheme cannot guarantee efficient utilization of RNs' resources since not many UEs can choose RNs as their serving nodes. In order to expand RN's coverage so that RN's resources can be effectively utilized by more UEs, a range-expansion based association scheme has been proposed [15]. Instead of attaching to the node which provides the strongest downlink signal strength, UEs can choose the node based on a biased received signal. With this scheme, RN's coverage can be effectively expanded. However, UEs located at RN's extended range will have weakened received signal, so that they might suffer strong interference from the neighboring high power BSs. To tackle these problems and more efficiently exploit RNs' resources, we introduce intra-cell cooperation in the heterogeneous relay networks.

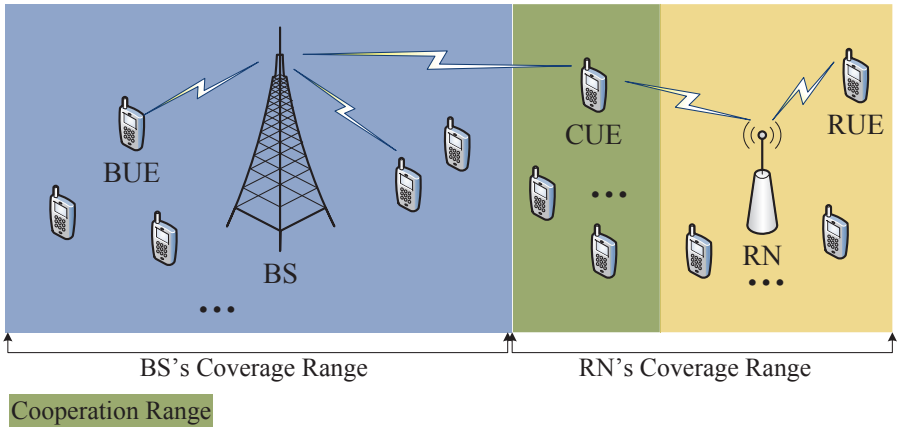

Fig. 2.1: Cooperative transmission in wireless heterogeneous network

As shown in Fig. 2.1, we classify UEs into three types. One type falls into BS's coverage range and is associated with BS. It receives transmission from the BS directly and is denoted as BUE. The second type and the third type are both associated with RNs. The second type, denoted as RUE, locates closely to a RN and directly receives transmissions from the RN and indirectly receives two-hop transmissions from the BS. The third type, denoted as CUE, locates at the extended coverage area of a RN and receives cooperative transmissions from the donor BS and the RN. The corresponding downlink received signal-to-interference-noise-ratio (SINR) for these UEs can be evaluated as follows. $\gamma^b_{i,0,k}$ is denoted as the SINR

value for UE k if it is a BUE and is served by BS in the ith sector. $\gamma^r_{i,j,k}$ is denoted as the SINR value for UE k if it is a RUE and served by jth RN in ith sector. $\gamma^{r,b}_{i,j,k}$ is denoted as the SINR value for UE k if it is a CUE and receives cooperative transmissions from jth RN and BS in the ith sector.

$$\gamma^b_{i,0,k} = \frac{P_m|h_{i,0,k}|^2}{N_0 + \sum_{i'=1,i'\neq i}^{N_c} |h_{i',0,k}|^2 P_m + \sum_{i'=1}^{N_c}\sum_{j'=1}^{N_r} |h_{i',j',k}|^2 P_r} \tag{2.1}$$

$$\gamma^r_{i,j,k} = \frac{P_r|h_{i,j,k}|^2}{N_0 + \sum_{\substack{i'=1\\(i',j')}}^{N_c}\sum_{\substack{j'=1\\\neq(i,j)}}^{N_r} |h_{i',j',k}|^2 P_r + \sum_{i'=1}^{N_c} |h_{i',0,k}|^2 P_m} \tag{2.2}$$

$$\gamma^{r,b}_{i,j,k} = \frac{P_m|h_{i,j,k}|^2 + P_r|h_{i,j,k}|^2}{N_0 + \sum_{\substack{i'=1\\(i',j')}}^{N_c}\sum_{\substack{j'=1\\\neq(i,j)}}^{N_r} |h_{i',j',k}|^2 P_r + \sum_{i'=1,i'\neq i}^{N_c} |h_{i',0,k}|^2 P_m} \tag{2.3}$$

$$i = 1, \cdots, N_c;\ j = 1, \cdots, N_r;\ k = 1, \cdots, N_u.$$

Here, $h_{i,0,k}$ represents the channel gain between BS in the ith sector and UE k, and $h_{i,j,k}$ represents the channel gain between the jth RN in the ith sector and UE k. N_0 is the variance of the additive white Gaussian noise (AWGN). Given the SINR value, the unit achievable data rate in terms of bit/s/Hz for each UE can be calculated using Shannon formula.

$$R^b_{i,0,k} = \log(1 + \gamma^b_{i,0,k}), \tag{2.4}$$

$$R^r_{i,j,k} = \log(1 + \gamma^r_{i,j,k}), \tag{2.5}$$

$$R^{r,b}_{i,j,k} = \log(1 + \gamma^{r,b}_{i,j,k}). \tag{2.6}$$

2.3 Problem Formulation

In this chapter, we want to design an optimal cooperation scheme that maximizes the long-term system throughput as well as ensure good user experience. The intra-cell

cooperation scheme works as follows.

For a UE associated with RN, if

- its γ is lower than a SINR threshold α, and
- the interference from its donor BS is greater than half of the received signal strength from RN,

then the UE will receive cooperative transmissions from both the serving RN and the donor BS. The second condition on interference makes sure that low SINR is interference-limited but not noise-limited, since the cooperative transmission in a strong noise-limited environment does not help too much.

Our goal is to derive the optimal threshold α for cooperative transmission. A too high α value will lead to an unnecessarily high number of UEs receiving cooperative transmissions while a low α value may leave certain UEs in bad SINR range receiving no cooperation. Both will lead to an undesirable system performance. The optimization problem consists of two tasks. The first task selects the best mobile association scheme. The second task optimizes the cooperative transmission. We can either jointly optimize these two tasks or optimize the cooperative transmission under a given mobile association scheme. In this chapter, we go for the latter design. In the conventional homogeneous networks, best-power based association scheme is widely used and is demonstrated to work well [16]. As stated in the previous section, it does not work well in the heterogeneous networks due to the disparity between BS and RN transmit powers. In order to let more UEs associate with RNs, a range-expansion based mobile association scheme has been proposed. This scheme uses a bias to compensate the power difference between BSs and RNs so that RN's coverage range can be expanded. The kth UE will choose the node $(i^*, j^*)_k$ (denote jth node in the ith sector. $(i, 0)$ represents the BS in ith sector) to associate with based on the following criterion

$$(i^*, j^*)_k = \arg \max_{i \in \{1,\cdots,N_c\}, j \in \{0,1,\cdots,N_r\}} (|h_{i,j,k}|^2 / \delta_{i,j}), \tag{2.7}$$

where $\delta_{i,0} = 1$ and $1 < \delta_{i,j} < (P_m/P_r), \text{for } j > 0$. $\delta_{i,j}$ value specifies the coverage of the macro- and micro-cells. A small $\delta_{i,j}$ leads to a large coverage region of the micro-cell while a large $\delta_{i,j}$ value leads to a small coverage region of the micro-cell. In extreme cases, $\delta_{i,j} = 1$ corresponds to path-loss based mobile association and $\delta_{i,j} = (P_m/P_r)$ corresponds to best-power based mobile association.

We use a decision variable $x_{i,0,k}$ to indicate the association status between the kth UE and the BS in ith sector. Specifically,

$$x^b_{i,0,k} = \begin{cases} 1 & \text{if } k\text{th UE is served by BS in the } i\text{th sector} \\ 0 & \text{otherwise.} \end{cases} \tag{2.8}$$

For UEs associated with RN, we further use $x^r_{i,j,k}$ and $x^{r,b}_{i,j,k}$ as the decision variables to denote the UEs served directly by jth RN and UEs jointly served by jth RN and the donor BS in the ith sector, respectively.

$$x^r_{i,j,k} = \begin{cases} 1 & \text{if } k\text{th UE is served by } j\text{th RN only} \\ 0 & \text{otherwise.} \end{cases} \tag{2.9}$$

$$x^{r,b}_{i,j,k} = \begin{cases} 1 & \text{if } k\text{th UE is served by } j\text{th RN and BS in the } i\text{th sector} \\ 0 & \text{otherwise.} \end{cases} \tag{2.10}$$

For UEs that are granted into the network, we have

$$\sum_{i=1}^{N_c} x^b_{i,0,k} + \sum_{i=1}^{N_c}\sum_{j=1}^{N_r}(x^r_{i,j,k} + x^{r,b}_{i,j,k}) = 1, \text{for all } k. \tag{2.11}$$

Next we formulate the optimal cooperation problem that aims to maximize the system throughput with user fairness. Several schemes have been proposed to address the fairness issue, such as the max-min fairness scheme proposed in [17], the proportional fairness scheme proposed in [18] and the competitive fairness scheme proposed in [19]. In this chapter, we use proportional fairness by defining the sum of log-scale throughput as the performance metric to optimize.

The optimization scheme decides the optimal α value based on time-averaged system-wide statistics, so that the updates on the α value will not need to occur per scheduling cycle. They will occur whenever the long-term statistics change. Denote the total frequency bands in BS and RN as C_i and $C_{i.j}$, respectively. For kth UE associated with the BS in the ith sector, the time-averaged allocated resources in the unit of sub-bands are denoted as $n^b_{i,0,k}$. Similarly, we denote $n^r_{i,j,k}$ as the time-averaged allocated resources for kth UE from the jth RN in the ith sector. For UE k which is jointly served by the ith BS and jth RN in the same cell, it has $n^{r,b}_{i,j,k}$ resources allocated from the BS and the RN. The optimization problem is formulated as follows.

$$\min -\sum_{i=1}^{N_c}\sum_{j=1}^{N_r}\sum_{k=1}^{N_u}\log(x^b_{i,0,k}n^b_{i,0,k}R^b_{i,0,k}+x^r_{i,j,k}n^r_{i,j,k}R^r_{i,j,k}+x^{r,b}_{i,j,k}n^{r,b}_{i,j,k}R^{r,b}_{i,j,k}) \tag{2.12}$$

subject to

$$\begin{aligned}
\sum_{k=1}^{N_u} x^b_{i,0,k}n^b_{i,0,k}+\sum_{j=1}^{N_r}\sum_{k=1}^{N_u} x^{r,b}_{i,j,k}n^{r,b}_{i,j,k} &\le C_i && (2.13)\\
\text{for } i &= 1,\cdots,N_c,\\
\sum_{k=1}^{N_u} x^r_{i,j,k}n^r_{i,j,k}+\sum_{k=1}^{N_u} x^{r,b}_{i,j,k}n^{r,b}_{i,j,k} &\le C_{i,j} && (2.14)\\
\text{for } i &= 1,\cdots,N_c; j=1,\cdots,N_r,\\
\sum_{i=1}^{N_c} x^b_{i,0,k}+\sum_{i=1}^{N_c}\sum_{j=1}^{N_r}(x^r_{i,j,k}+x^{r,b}_{i,j,k}) &= 1, && (2.15)\\
x^b_{i,0,k}n^b_{i,0,k},\ x^r_{i,j,k}n^r_{i,j,k},\ x^{r,b}_{i,j,k}n^{r,b}_{i,j,k} &\ge 0 && (2.16)\\
\text{for } i &= 1,\cdots,N_c; j=1,\cdots,N_r; k=1,\cdots,N_u,
\end{aligned}$$

where $n^b_{i,0,k}R^b_{i,0,k}$ is the long-term time-averaged throughput for the kth UE when associated with the BS in the ith sector. $n^r_{i,j,k}R^r_{i,j,k}$ is the long-term time-averaged throughput for the kth UE associated with jth RN in the ith sector without cooperation. $n^{r,b}_{i,j,k}R^{r,b}_{i,j,k}$ is the long-term time-averaged throughput for the kth UE associated with jth RN in the ith sector with cooperation. $R^b_{i,0,k}$, $R^r_{i,j,k}$ and $R^{r,b}_{i,j,k}$ are defined in (2.4)-(2.6). $n^b_{i,0,k}$, $n^r_{i,j,k}$ and $n^{r,b}_{i,j,k}$ can be non-integers as they represent time-averaged values. The log-scale throughput objective

function can achieve a good balance between throughput maximization and fairness since any increase of an already large throughput for an individual UE will only lead to a marginal increase on the objective function. Constraint (2.13) and (2.14) regulate the usage of the resources at the BSs and RNs. Constraint (2.15) makes sure a granted UE can only be one of the three types defined before.

2.4 Optimal Cooperative Transmission Algorithm

Our goal is to solve the optimal α from (2.12)-(2.16). Actually α does not directly show up in the optimal problem defined above. The primal problem is non-convex and it is difficult to derive its optimal solution. However, if we fix the value of α, then $x^b_{i,0,k}$, $x^r_{i,j,k}$, $x^{r,b}_{i,j,k}$ and $R^b_{i,0,k}$, $R^r_{i,j,k}$, $R^{r,b}_{i,j,k}$ can all be decided based on the given mobile association scheme and α value. The primal optimization problem becomes convex and it reduces to an optimization problem with variables $n^m_{i,j,k}$, for $m = (b, r, (r, b))$. Therefore, we propose a two-loop procedure to solve the primal optimization problem. α value is optimized in the outer loop using a brute-force search. In the inner loop, given the α value specified in the outer loop, the original optimization problem becomes a constraint convex optimization problem with variables $n^m_{i,j,k}$. The optimal solutions can be found by solving the corresponding dual problem. In the following, we present the details of the optimization procedure.

Introducing Lagrange multipliers λ^b_i, $\lambda^r_{i,j}$ and λ^m_k, for $m = (b, r, (r, b))$ (all are non-negative), the Lagrange function can be formed as

$$\begin{aligned}
\mathcal{L}(n^m_{i,j,k}, \boldsymbol{\lambda}) \;=\; & -\sum_{i=1}^{N_c}\sum_{j=1}^{N_r}\sum_{k=1}^{N_u} \log(x^b_{i,0,k} n^b_{i,0,k} R^b_{i,0,k} + x^r_{i,j,k} n^r_{i,j,k} R^r_{i,j,k} + x^{r,b}_{i,j,k} n^{r,b}_{i,j,k} R^{r,b}_{i,j,k}) \\
& + \sum_{i=1}^{N_c} \lambda^b_i \Big(\sum_{k=1}^{N_u} x^b_{i,0,k} n^b_{i,0,k} + \sum_{j=1}^{N_r}\sum_{k=1}^{N_u} x^{r,b}_{i,j,k} n^{r,b}_{i,j,k} - C_i \Big) \\
& + \sum_{i=1}^{N_c}\sum_{j=1}^{N_r} \lambda^r_{i,j} \Big(\sum_{k=1}^{N_u} x^r_{i,j,k} n^r_{i,j,k} + \sum_{k=1}^{N_u} x^{r,b}_{i,j,k} n^{r,b}_{i,j,k} - C_{i,j} \Big) \\
& - \sum_{i=1}^{N_c}\sum_{j=1}^{N_r}\sum_{k=1}^{N_u} (\lambda^b_k x^b_{i,0,k} n^b_{i,0,k} + \lambda^r_k x^r_{i,j,k} n^r_{i,j,k} + \lambda^{r,b}_k x^{r,b}_{i,j,k} n^{r,b}_{i,j,k}). \qquad (2.17)
\end{aligned}$$

The corresponding dual function and dual problem are:

$$g(\boldsymbol{\lambda}) = \inf_{n^m_{i,j,k}} \mathcal{L}(n^m_{i,j,k}, \boldsymbol{\lambda}) \tag{2.18}$$
$$\lambda^b_i,\ \lambda^r_{i,j},\ \lambda^m_k \in \boldsymbol{\lambda}, \text{for } m = (b, r, (r, b))$$

and

$$\max\ g(\boldsymbol{\lambda}) \tag{2.19}$$

subject to

$$\lambda^b_i \geq 0,\ \lambda^r_{i,j} \geq 0,\ \lambda^m_k \geq 0.$$

Notice that the dual function (2.18) is always concave and the dual problem (2.19) is always convex [20]. As mentioned earlier, when α value is fixed, the primal function is convex and the constraints (2.13)-(2.15) are linear, so that the optimization problem satisfies Slater's condition, and the strong duality holds [21]. Thus, the optimal solutions for primal problem can be obtained from the dual problem. If we denote the primal problem as $f_0(n^m_{i,j,k}),\ \text{for } m = (b, r, (r, b))$, we have the following primal-dual optimality:

$$f_0(n^{m*}_{i,j,k}) = g(\boldsymbol{\lambda}^*) = \inf_{n^m_{i,j,k}} \mathcal{L}(n^m_{i,j,k}, \boldsymbol{\lambda}^*). \tag{2.20}$$

2.4.1 Optimal $n^b_{i,0,k}$, $n^r_{i,j,k}$ and $n^{r,b}_{i,j,k}$

In order to solve the dual optimization problem, we need to derive the expression of the dual function $g(\boldsymbol{\lambda})$ at first. As the dual function is a point-wise minimum of a family of linear functions of the Lagrange multipliers, we could find out the optimal $n^m_{i,j,k}$'s that minimize the $\mathcal{L}(n^m_{i,j,k}, \boldsymbol{\lambda^*}),\ \text{for } m = (b, r, (r, b))$. The optimal $n^m_{i,j,k}$'s can be found by setting the gradient of $\mathcal{L}(n^m_{i,j,k}, \boldsymbol{\lambda^*})$ with respect to $n^m_{i,j,k}$ equal to zero:

$$\frac{\partial \mathcal{L}(n^m_{i,j,k}, \boldsymbol{\lambda^*})}{\partial n^m_{i,j,k}} = 0 \text{ for } m = (b, r, (r, b)), \tag{2.21}$$

then we can obtain

$$n_{i,0,k}^{b*} = \frac{1}{\lambda_i^{b*} x_{i,0,k}^b - \lambda_k^{b*} x_{i,0,k}^b}, \tag{2.22}$$

$$n_{i,j,k}^{r*} = \frac{1}{\lambda_{i,j}^{r*} x_{i,j,k}^r - \lambda_k^{r*} x_{i,j,k}^r}, \tag{2.23}$$

and

$$n_{i,j,k}^{(r,b)*} = \frac{1}{\lambda_i^{b*} x_{i,j,k}^{r,b} + \lambda_{i,j}^{r*} x_{i,j,k}^{r,b} - \lambda_k^{(r,b)*} x_{i,j,k}^{r,b}}. \tag{2.24}$$

2.4.2 Optimal Values for Lagrange Multipliers λ_i^b, $\lambda_{i,j}^r$ and λ_k^m

Without loss of generality, substituting (2.22)-(2.24) into (2.17), we can get the dual function $g(\boldsymbol{\lambda}^*)$. Because of the concavity of the dual function (2.18), we can use gradient-descent method to search the optimal λ_i^{b*}, $\lambda_{i,j}^{r*}$ and λ_k^{m*}, for $m = (b, r, (r, b))$. By taking the gradient of $g(\boldsymbol{\lambda})$ with respect to λ_i^b, $\lambda_{i,j}^r$ and λ_k^m, we can obtain

$$\Delta\lambda_i^b(t) = \sum_{j=1}^{N_r}\sum_{k=1}^{N_u} \frac{x_{i,j,k}^{r,b}}{\lambda_i^b x_{i,j,k}^{r,b} + \lambda_{i,j}^r x_{i,j,k}^{r,b} - \lambda_k^{r,b} x_{i,j,k}^{r,b}} + \sum_{k=1}^{N_u} \frac{x_{i,0,k}^b}{\lambda_i^b x_{i,0,k}^b - \lambda_k^b x_{i,0,k}^b} - C_i, \tag{2.25}$$

$$\Delta\lambda_{i,j}^r(t) = \sum_{k=1}^{N_u} \frac{x_{i,j,k}^{r,b}}{\lambda_i^b x_{i,j,k}^{r,b} + \lambda_{i,j}^r x_{i,j,k}^{r,b} - \lambda_k^{r,b} x_{i,j,k}^{r,b}} + \sum_{k=1}^{N_u} \frac{x_{i,j,k}^r}{\lambda_{i,j}^r x_{i,j,k}^r - \lambda_k^r x_{i,j,k}^r} - C_{i,j}, \tag{2.26}$$

$$\Delta\lambda_k^b(t) = \frac{x_{i,0,k}^b}{\lambda_i^b x_{i,0,k}^b - \lambda_k^b x_{i,0,k}^b}, \tag{2.27}$$

$$\Delta\lambda_k^r(t) = \frac{x_{i,j,k}^r}{\lambda_{i,j}^r x_{i,j,k}^r - \lambda_k^r x_{i,j,k}^r}, \tag{2.28}$$

$$\Delta\lambda_k^{r,b}(t) = \frac{x_{i,j,k}^{r,b}}{\lambda_i^b x_{i,j,k}^{r,b} + \lambda_{i,j}^r x_{i,j,k}^{r,b} - \lambda_k^{r,b} x_{i,j,k}^{r,b}}. \tag{2.29}$$

We update λ_i^b, $\lambda_{i,j}^r$, and λ_k^m simultaneously along the directions

$$\lambda_i^b(t+1) = \lambda_i^b(t) + \mu\Delta\lambda_i^b(t), \tag{2.30}$$

$$\lambda_{i,j}^r(t+1) = \lambda_{i,j}^r(t) + \mu\Delta\lambda_{i,j}^r(t), \tag{2.31}$$

$$\lambda_k^m(t+1) = \lambda_k^m(t) + \mu\Delta\lambda_k^m(t), \tag{2.32}$$

where $m = (b, r, (r, b))$, μ is the step size for each update. If $|\Delta\lambda_i^b(t)| \leq \epsilon$, or $|\Delta\lambda_{i,j}^r(t)| \leq \epsilon$, or $|\Delta\lambda_k^m(t)| \leq \epsilon$ (ϵ is a very small positive value), we claim that λ_i^b, or $\lambda_{i,j}^r(t)$ or $\lambda_k^m(t)$ converges. Once we obtain the optimal Lagrange multipliers, we can calculate optimal $n_{i,0,k}^b$, $n_{i,j,k}^r$ and $n_{i,j,k}^{r,b}$ by substituting the optimal Lagrange multipliers into (2.22)-(2.24).

2.4.3 Summary of Optimization Procedure

A summary of the proposed two-loop optimization procedure is given as follows.

Outer-loop

Step-1: Set up a fixed bias value δ and determine the association status for each UE.

Step-2: Given a SINR threshold α, for the UE associated with RN, we further decide if a cooperative transmission is needed or not based on the following.

(1) SINR is lower than α;

(2) The interference power from neighboring BS $P_I \geq 0.5P_{\text{re}}$ where P_{re} is the received downlink power from UE's serving RN;

Step-3: Based on Step-1 and Step-2, we categorize UEs into three groups: UEs served by BSs, UEs served by RNs and UEs served by cooperative transmissions. Then step to inner-loop.

Inner-loop

Step-4: Initialize Lagrange multipliers $\lambda_i^b(0)$, $\lambda_{i,j}^r(0)$ and $\lambda_k^m(0)$, for $m = (b, r, (r, b))$.

Step-5: In each iteration, we can compute the $\Delta\lambda_i^b(t)$, $\Delta\lambda_{i,j}^r(t)$ and $\Delta\lambda_k^m(t)$ using (2.25)-(2.29). Then update λ_i^b, $\lambda_{i,j}^r$, and λ_k^m through (2.30)-(2.32).

Step-6: Repeat Step-5 until the updates on $\lambda_i^b(t)$, $\lambda_{i,j}^r(t)$ and $\lambda_k^m(t)$ converge. Then substituting the optimal Lagrange multipliers into (2.22)-(2.24) and (2.12), we can obtain the optimal $n_{i,0,k}^b$, $n_{i,j,k}^r$ and $n_{i,j,k}^{r,b}$, and the optimal objective function value.

End(Inner-loop)

Step-7: Update the SINR threshold α as $\alpha(\tau + 1) = \alpha(\tau) + \Delta\alpha$. Repeat Step-2 to Step-6.

End(Outer-loop)

Step-8: Find the global optimal solution (α^*, $n_{i,0,k}^{b*}$, $n_{i,j,k}^{r*}$, $n_{i,j,k}^{(r,b)*}$) that gives the highest objective function in the above two loop search.

For additional clarity, the two-loop optimization procedure is summarized in Fig. 2.2.

2.5 Performance Evaluation

We simulate a cellular network with a 19-cell 3-sector three-ring hexagonal cell structure with a cell radius at 2 km. Four RNs are uniformly deployed in each sector. Simulation setup follows the guidelines described in 3GPP technical reports [22]. Transmit power of a BS is 46dBm (40W) and transmit power of a RN is 30dBm (1W). UEs are uniformly distributed in the network with an average of 200 UEs per cell.

The first simulation compares intra-cell CT scheme with inter-cell CT scheme. Intra-cell CT is formed between RN and BS in the same cell to minimize the data exchange and signaling overhead. In the inter-cell CT, we allow the cooperation formed between RN and any BS that causes the strongest interference to the UE. In Fig. 2.3, we set the bias value $\delta = 0$ dB and plot the objective function defined in (2.12) for different α values (maximum problem is equivalent to negative minimum problem). The log-scale system throughput achieves the maximum value for intra-/inter-cell CTs at $\alpha^* = -9.214$ dB and $\alpha^* = -8.7$ dB, respectively. When α exceeds the optimal value, more RN-associated UEs will receive cooperative transmissions. The system throughput decreases since the UE throughput gained from cooperative transmissions does not make up the double resources consumed by the cooperative transmissions from BS and RN. On the other hand, when the selected α is below the optimal value, fewer UEs will use cooperative transmissions, including some UEs at low SINR. Extensive RN radio resources are consumed to support the low SINR UEs so that the overall system log throughput actually goes down. A very low α, e.g., -25 dB, practically represents a case without intra-cell CT. Compared to the system without intra-cell CT, intra-cell CT with optimal α can not only achieve 6% gain on the log-scale throughput but also results in a much better SINR improvement, as shown later in Fig. 2.6. In addition, by comparing intra-cell CT and inter-cell CT, we find inter-cell

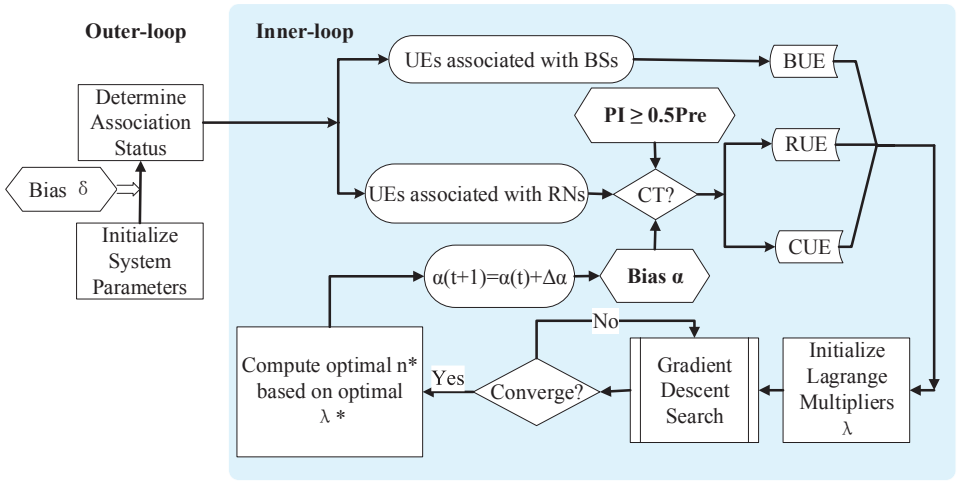

Fig. 2.2: Two-loop optimization procedure

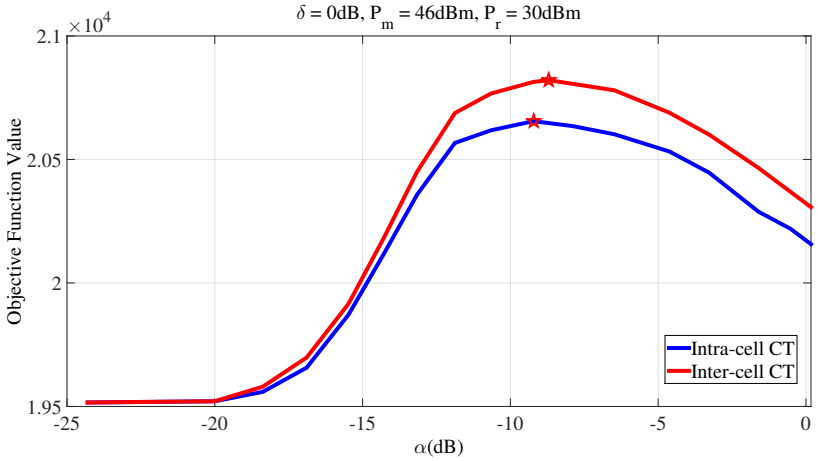

Fig. 2.3: Intra-cell CT vs. Inter-cell CT: $P_m = 46$dBm, $P_r = 30$dBm, $\delta = 0$ dB

CT always outperforms intra-cell CT if using the same α. The results are aligned with the expectation. The inter-cell CT selects the BS who contributes the strongest interference to the UE to form cooperative transmission. As a result, the UE's SINR can be better improved than the intra-cell CT. However, the cooperative BS in the inter-cell CT could locate in another cell and there may be no direct connection between the RN and the cooperative BS. The inter-cell CT incurs a much higher data exchange overhead on the backhaul, a much higher implementation complexity and a much longer scheduling delay. So in a distributed BS deployment scenario, inter-cell CT will have limited applications.

We further investigate the impact of mobile association strategies on the intra-cell CT

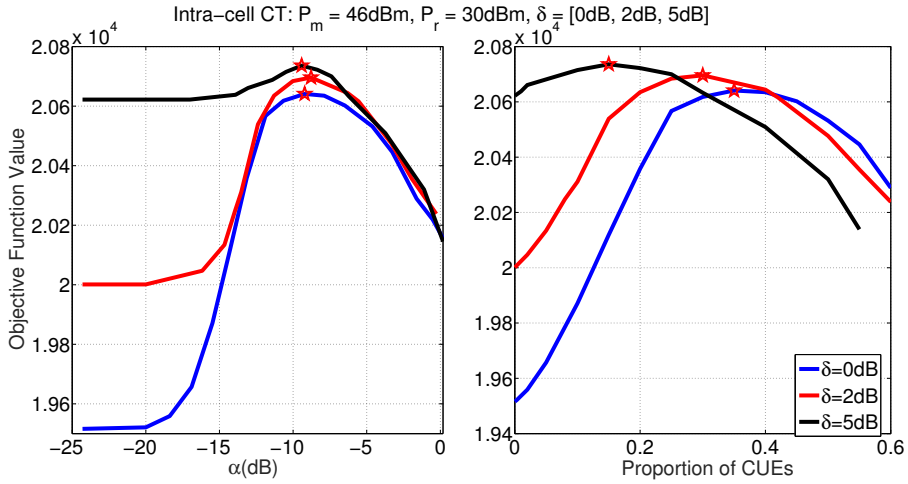

Fig. 2.4: Intra-cell CT at different mobile association bias values

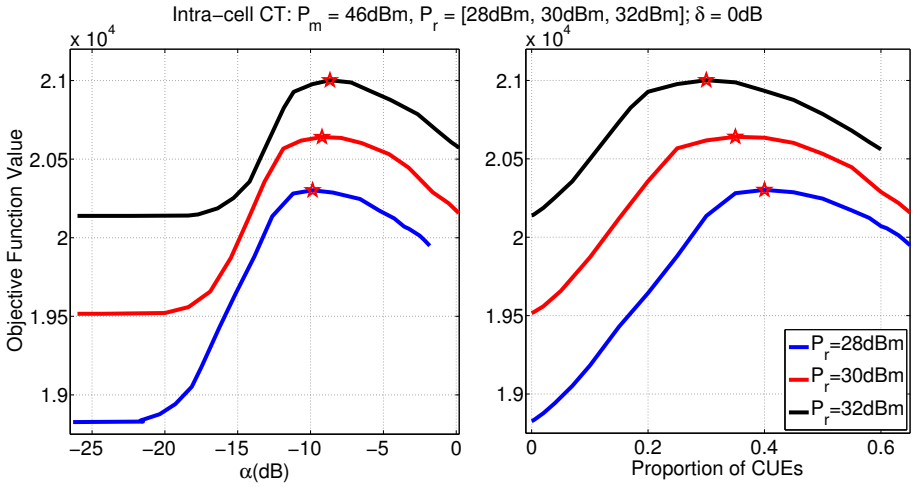

Fig. 2.5: Intra-cell CT at different RN transmit powers

performance. In Fig. 2.4, we compare the objective function values at different δ values. Note that $\delta = 0$ dB corresponds to the path-loss based mobile association and $\delta = 16$ dB corresponds to the best-power based mobile association. It is observed that a higher δ value will lead to a lower percentage of CUEs. If the δ increases from 0 dB to 5 dB, the portion of CUEs in the total UEs decreases from 35% to 15%. With a higher δ value, fewer UEs will be associated with RNs. In another word, the coverage range of RNs will be smaller given a higher δ value. Thus the number of UEs that are far away from RNs and exposed to strong interferences from nearby high power BSs will reduce. Therefore, fewer UEs need

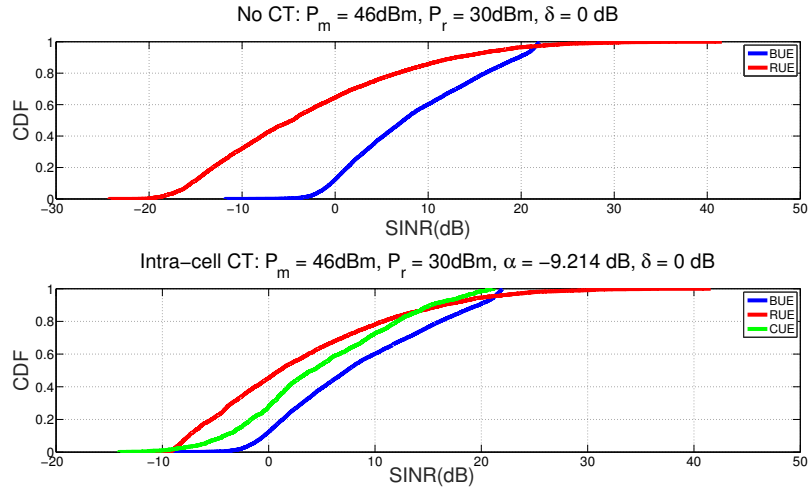

Fig. 2.6: Non-CT vs. Intra-cell CT: CDF of UEs' SINR

cooperative transmissions.

The next simulation shows how RN transmit power impacts the intra-cell CT performance. If RN's transmit power increases from 28 dBm to 32 dBm, we can observe from Fig. 2.5 that the percentage of CUEs decreases from 40% to 30%. When RN's transmit power increases, more UEs in the RN's coverage can receive good SINRs. Thus, fewer UEs will need cooperative transmissions.

Fig. 2.6 compares the SINR distribution between the cases with and without intra-cell CT. Both cases use range-expansion based mobile association with $\delta = 0$ dB. Without intra-cell CT, more than 30% UEs associated with RN have an SINR below -10 dB. By choosing $\alpha^* = -9.214$ dB and deploying the intra-cell CT, the SINR distribution of the UEs at RN's coverage range is improved by about 10 dB. Only 1% UEs have an SINR below -10 dB. UEs which suffer strong interference and thus receive poor received downlink SINR can leverage intra-cell CT to improve the performance tremendously.

2.6 Chapter Summary

In this chapter, we investigated the downlink intra-cell cooperative transmission in the heterogeneous networks and developed an optimal cooperation scheme to achieve both throughput maximization and user fairness. The scheme is optimized by selecting the best

SINR threshold to form intra-cell cooperation. The optimization is based on long-term time-averaged system information and only needs to updated pseudo-dynamically. Simulation results showed that the cooperative transmission can greatly improve the network performance in a heterogeneous network.

Chapter 3
Optimal CoMP with Precoding in Wireless Heterogeneous Networks

3.1 Introduction

In the previous chapter, we explored the advantages of cooperative transmission in wireless heterogeneous networks. With proper mobile association scheme and node cooperation, we can greatly improve the system-wise spectral efficiency and user experience. Thereby, coordinated multi-point processing (CoMP) is considered as an important approach to improve the performance for cell edge users. Its application in a wireless heterogeneous network resource allocation has also been addressed in [23–25]. Considering a heterogeneous network in which both mBSs and pBSs are deployed, the mBSs and pBSs can coordinate on scheduling and data transmission among adjacent cells to improve the coverage and cell edge spectral quality [3, 26]. In addition, precoding applies an appropriate weight to the signal emitted from each of the transmitting antennas such that the signal power is maximized on the receiving side. Thus, precoding scheme can be combined with CoMP technique to further improve the cell edge performance and achieve substantial capacity gains. In this chapter, we employ Tomlinson-Harashima precoding (THP) [27] with CoMP, so that mBS and pBSs can serve multiple UEs simultaneously at the cell edge and achieve more efficient utilization of radio resources.

3.2 Network Model and Precoder Design

Without loss of generality, we consider a downlink communication system in a wireless heterogeneous network shown in Fig. 1.1. Each cell is divided into multiple sectors, and in each sector, one mBS and multiple pBSs are deployed. We denote the total number of mBSs as N_c and the number of pBSs per sector as N_r. Then, the total number of pBSs

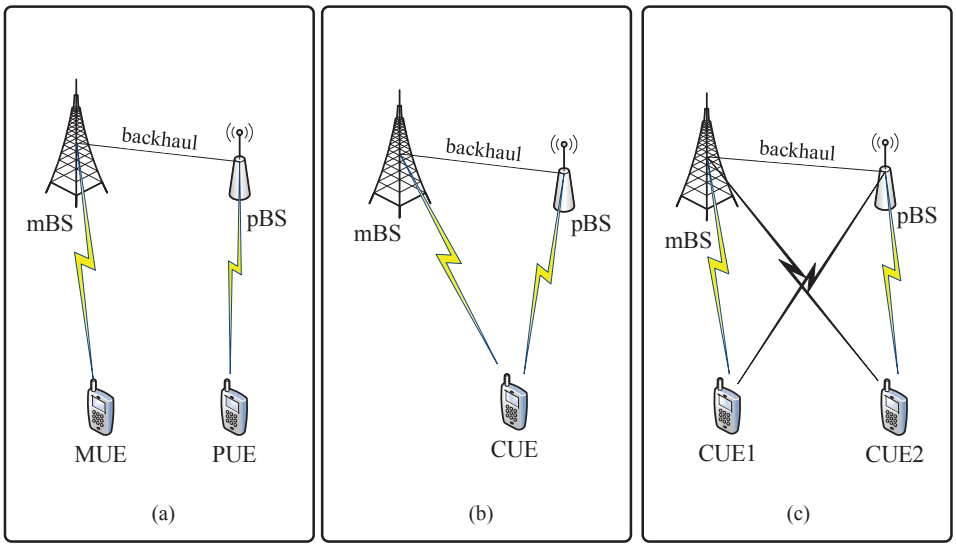

Fig. 3.1: Service modes: (a) No CoMP; (b) CoMP without precoding; (c) CoMP with precoding

in the network is $N_p = N_c \times N_r$. An mBS has a high transmit power P_m, thereby it is used to provide blanket coverage and seamless mobility. A pBS usually has a much lower transmit power P_p that generates a small footprint to support cell edge users and to boost local capacity. N_u UEs are uniformly distributed in the network. UEs receive the service based on three different association types. A UE falling into an mBS's coverage range is associated with the mBS and is denoted as macro-UE (MUE) (Fig. 3.1-a). The second type, denoted as pico-UE (PUE), locates closely to a pBS and is associated with the pBS. It directly receives transmission from a pBS and indirectly receives two-hop transmissions from an mBS (Fig. 3.1-a). The third type, denoted as cooperative UE (CUE), locates at the extended coverage area of a pBS and is associated with the pBS. But it receives cooperative transmissions from mBS and pBS (Fig. 3.1-b, c). For scheduling purpose, we divide the total frequency band into F resource blocks (RBs) and each UE can be assigned with an integer number of RBs at time t. Each node (mBS or pBS) is equipped with only one antenna. However, for the CUEs that receive cooperative transmissions or joint processing from both mBS and pBS, we can view the two transmitting nodes as two antennas in a co-located MIMO system, thus forming a network MIMO for the CUEs. In order to maximize the sum

throughput in a network MIMO system, we allow two CUEs to simultaneously receive from the same mBS and pBS (Fig. 3.1-c) by applying THP precoding algorithm [27], which can eliminate inter-user interference between the two CUEs that receive from the same mBS and pBS at the same frequency band. We first formulate a 2×2 channel matrix $\mathbf{H}$

$$\mathbf{H} = \begin{bmatrix} \sqrt{P_m}h_{1,1} & \sqrt{P_p}h_{1,2} \\ \sqrt{P_m}h_{2,1} & \sqrt{P_p}h_{2,2} \end{bmatrix}. \tag{3.1}$$

Here, $h_{m,n}$ is denoted as the channel gain. Based on [27], the precoding matrix is designed as

$$\mathbf{W} = \mathbf{F}\mathbf{B}^{-\mathbf{1}}\mathbf{J}. \tag{3.2}$$

in which matrices $\mathbf{F}$, $\mathbf{B}$ and $\mathbf{J}$ are given by

$$\begin{aligned} &\mathbf{H}^* = \mathbf{Q}\mathbf{R}^*; \quad \mathbf{B} = \mathbf{G}\mathbf{R}; \quad \mathbf{F} = \mathbf{Q}; \\ &\mathbf{G} = \mathrm{diag}[\frac{1}{|r_{11}|} \quad \frac{1}{|r_{22}|}]; \quad \mathbf{J} = \mathrm{diag}[\frac{r_{11}}{|r_{11}|} \quad \frac{r_{22}}{|r_{22}|}], \end{aligned} \tag{3.3}$$

where $\mathbf{Q}$ is a unitary matrix or semi-unitary matrix and $\mathbf{R}$ is a lower triangle matrix. $\mathbf{H}^*$ is the conjugate transpose of $\mathbf{H}$ and $\mathbf{R}^*$ is the conjugate transpose of $\mathbf{R}$. r_{kk} is the diagonal element of $\mathbf{R}$ in the kth row and $\mathbf{J}$ is a local phase adjustment matrix which is used to combine the channel gains coherently at UE.

The received signal $\mathbf{y} = [y_1 \quad y_2]^T$ at the CUE is

$$\mathbf{y} = \mathbf{H}\mathbf{W}\mathbf{x} + \mathbf{n} = \mathbf{H}\mathbf{F}\mathbf{B}^{-\mathbf{1}}\mathbf{J}\mathbf{x} + \mathbf{n} = \begin{bmatrix} r_{11} & 0 \\ 0 & r_{22} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \begin{bmatrix} n_1 \\ n_2 \end{bmatrix}. \tag{3.4}$$

The corresponding downlink received SINR for MUE, PUE and CUE can be evaluated as follows.

$$\gamma^f_{i,0,k}(t) = \frac{P_m|h^f_{i,0,k}(t)|^2}{N_0 + \sum_{i'=1,i'\neq i}^{N_c}|h^f_{i,0,k'}(t)|^2P_m + \sum_{i'=1}^{N_c}\sum_{j'=1}^{N_r}|h^f_{k,j',i'}(t)|^2P_p} \tag{3.5}$$

$$\gamma^f_{i,j,k}(t) = \frac{P_p|h^f_{i,j,k}(t)|^2}{N_0 + \sum_{\substack{i'=1\\(i',j')}}^{N_c}\sum_{\substack{j'=1\\\neq(i,j)}}^{N_r}|h^f_{i',j',k}(t)|^2P_p + \sum_{i'=1}^{N_c}|h^f_{i,0,k'}(t)|^2P_m} \tag{3.6}$$

$$\gamma^{c,f}_{i,j,k}(t) = \frac{|r^f_{11(22)}(t)|^2}{N_0 + \sum_{\substack{i'=1\\(i',j')}}^{N_c}\sum_{\substack{j'=1\\\neq(i,j)}}^{N_r}|h^f_{i',j',k}(t)|^2P_p + \sum_{i'=1,i'\neq i}^{N_c}|h^f_{i,0,k'}(t)|^2P_m} \tag{3.7}$$

$$i = 1,\cdots,N_c;\ j = 1,\cdots,N_r;\ k = 1,\cdots,N_u.$$

Here, $h^f_{i,0,k}(t)$ is the channel gain of the fth RB at time t between the ith mBS and the kth UE, and $h^f_{i,j,k}(t)$ is the channel gain of the fth RB at time t between the jth pBS in the ith sector and the kth UE. They both include long-term path loss, shadowing and short term fading due to multipath and mobility. r^f_{11} is the equivalent channel gain of the fth RB between the ith mBS and the CUE1. r^f_{22} is the equivalent channel gain of the fth RB between the jth pBS in the ith sector and the CUE2. N_0 is the variance of the additive white Gaussian noise. Given SINR, the unit achievable data rate in terms of bit/s/Hz for each UE can be calculated using Shannon formula.

$$R^b_a = \log(1+\gamma^b_a) \quad \text{for } a = ((i,0,k),(i,j,k)) \text{ and } b = (f,(c,f)). \tag{3.8}$$

3.3 Problem Formulation

Our objective is to optimize the network long-term spectrum efficiency and service fairness. Towards that end, we need to:

(1) properly decide the association for each UE;

(2) properly allocate RBs to the UEs at each scheduling cycle.

In order to expand the pBS's coverage range so that pBS's resource can serve more UEs, we also apply bias-based range-expansion mobile association scheme, which is explicitly described in the previous chapter.

We denote $x_{i,0,k}$ as the decision variable to indicate the association status between the kth UE and the ith mBS. Specifically,

$$x_{i,0,k} = \begin{cases} 1 & \text{if } k\text{th UE is associated with } i\text{th mBS} \\ 0 & \text{otherwise.} \end{cases} \tag{3.9}$$

$x_{i,j,k}$ is similarly defined for UEs associated with pBSs. Each UE can only attach to one BS, i.e., $\sum_{i=1}^{N_c} \sum_{j=0}^{N_r} x_{i,j,k} \leq 1, \ \forall k$. Furthermore, $x_{i,j,k}^{c,f}(t)$ is used to denote if CoMP is used or not at each scheduling cycle t. $x_{i,j,k}^{c,f}(t) = 1$ indicates that UE k is jointly served on RB f by jth pBS and the donor mBS in the ith sector while $x_{i,j,k}^{c,f}(t) = 0$ indicates UE k receives transmission only from pBS on RB f. Unlike $x_{i,j,k}$ or $x_{i,0,k}$, which is decided during the mobile association stage, $x_{i,j,k}^{c,f}(t)$ is decided at t based on the instantaneous channel state.

We denote $\mathcal{K}_{i,0}$ as the set of MUEs associated with ith mBS, $\mathcal{K}_{i,j}$ as the set of PUEs associated with jth pBS in the ith sector. At t, we also decide the set of CUEs $\mathcal{K}_{i,j}^{c,f}(t)$ that are associated with pBS j in sector i and are the candidates for joint processing by mBS and pBS on RB f.

$$\mathcal{K}_{i,j}^{c,f}(t) = \{k \in \mathcal{K}_{i,j} | \gamma_{i,j,k}^{f}(t) < \alpha\}, \tag{3.10}$$

where α is the SINR threshold that decides the CoMP set.

In order to formulate the scheduling problem, we introduce the following variables. $n_{i,0,k}^{f}(t) = 1(\text{or } 0)$ means that the fth RB is (is not) assigned to the kth MUE in the ith sector at time t, $n_{i,j,k}^{f}(t) = 1(\text{or } 0)$ indicates that the fth RB is (is not) assigned to the kth PUE at the jth pBS in the ith sector at time t, $n_{i,j,k}^{c,f}(t) = 1(\text{or } 0)$ indicates that the fth RB is (is not) assigned to the kth CUE served by the jth pBS and the ith sector at time t.

We use proportional fairness as the performance metric to ensure a good trade-off

between spectrum efficiency and fairness. The optimization problem with a long-term proportional fair resource allocation is thus formulated as

$$\mathbf{P}_1: \quad U(\mathbf{R}(t)) := \max_{\mathbf{n}(t)} \sum_k \log(R_k(t)) \tag{3.11}$$
$$\text{for } \mathbf{n}(t) = \left\{ n^f_{i,0,k}(t), n^f_{i,j,k}(t), n^{c,f}_{i,j,k}(t) \right\}$$

subject to

$$\sum_{k=1}^{N_u} x_{i,0,k} n^f_{i,0,k}(t) + \sum_{j=1}^{N_r} \sum_{k=1}^{N_u} x_{i,j,k} x^{c,f}_{i,j,k}(t) n^{c,f}_{i,j,k}(t) \leq 1 \tag{3.12}$$
$$\text{for } i = 1, \cdots, N_c, f = 1, \cdots, F$$
$$\sum_{k=1}^{N_u} x_{i,j,k}(1 - x^{c,f}_{i,j,k}(t)) n^f_{i,j,k}(t) + \sum_{k=1}^{N_u} x_{i,j,k} x^{c,f}_{i,j,k}(t) n^{c,f}_{i,j,k}(t) \leq 1 \tag{3.13}$$
$$\text{for } i = 1, \cdots, N_c, \ j = 1, \cdots, N_r, f = 1, \cdots, F$$

$$n^f_{i,j,k}(t) = 0 \text{ or } 1 \tag{3.14}$$

$$n^{c,f}_{i,j,k}(t) = 0 \text{ or } 1, \ \forall i, j, k, f, \tag{3.15}$$

where

$$R_k(t) = \frac{1}{T_c} \sum_{\tau=t-T_c+1}^{t} S_k(\tau), \tag{3.16}$$

T_c is the size of time window for moving average, and S_k is the moving average system throughput, which is expressed as

$$S_k(\tau) = \sum_{f=1}^{F} \sum_{i=1}^{N_c} \sum_{j=0}^{N_r} \left(x_{i,j,k}(1 - x^{c,f}_{i,j,k}(\tau)) R^f_{i,j,k}(\tau) n^f_{i,j,k}(\tau) + x_{i,j,k} x^{c,f}_{i,j,k}(\tau) R^{c,f}_{i,j,k}(\tau) n^{c,f}_{i,j,k}(\tau) \right), \tag{3.17}$$

where we set $x^{c,f}_{i,0,k}(\tau) = 0$ for notational consistency.

By solving the $n^f_{i,j,k}(t)$ and $n^{c,f}_{i,j,k}(t)$ values in $\mathbf{P}_1$, we can find the allocated UE for each RB. The problem is a 0-1 Knapsack problem and is NP-hard. We first relax the domain of the integers $n^f_{i,j,k}(t)$ and $n^{c,f}_{i,j,k}(t)$ into real number, i.e., $n^f_{i,j,k}(t) \in [0, 1]$ and $n^{c,f}_{i,j,k}(t) \in [0, 1]$. By doing so, (3.12) is the resource constraint for each RB at the mBS. The first term

in (3.12) represents the portion of RB f used by MUEs, and the second term in (3.12) computes the portion of the RB f used at the mBS to serve the CUEs. Similarly, (3.13) gives the resource constraint for each RB at the pBS. The first term represents the portion of RB f used by PUEs with no CoMP while the second term represents the portion of RB f used by the PUEs with CoMP, i.e., CUEs. As a multicarrier proportional fair scheduling problem, it is hard to find the optimal solution of $\mathbf{P}_1$ directly [28]. Considering practical implementation, we apply the gradient descent based scheduling algorithm in [29], which proved that the gradient descent based scheduling algorithm asymptotically converges to the optimal solution. In the next section, based on the gradient descent based scheduling algorithm, we show how to optimally allocate resources in such a heterogeneous network.

3.4 An Asymptotically Optimal Radio Resource Scheduling Scheme

Using the gradient descent based scheduling framework, the system parameters are chosen to maximize the drift of the objective function at each subframe, given as

$$\begin{aligned} U(\mathbf{R}(t+1)) - U(\mathbf{R}(t)) &= \sum_{k=1}^{N_u} \left(\log\Big(R_k(t) + \epsilon\big(S_k(t+1) - S_k(t-T_c+1)\big)\Big) - \log\big(R_k(t)\big)\right) \\ &= \sum_{k=1}^{N_u} \frac{1}{R_k(t)} S_k(t+1)\epsilon - \sum_{k=1}^{N_u} \frac{1}{R_k(t)} S_k(t-T_c+1)\epsilon + O(\epsilon^2), \quad (3.18) \end{aligned}$$

where $\epsilon = 1/T_c$ and the second equality is obtained using first order Taylor expansion. Since only the first term in (3.18) depends on future decisions and constraints (3.12)-(3.15) are set on a per RB basis, we can formulate the gradient descent based scheduling problem for each RB f as $\mathbf{P}_2$:

$$\max_{\mathbf{n}(t)} \sum_{k=1}^{N_u} \sum_{i=1}^{N_c} \sum_{j=0}^{N_r} \frac{\Big[x_{i,j,k}(1 - x_{i,j,k}^{c,f}(t)) R_{i,j,k}^{f}(t) n_{i,j,k}^{f}(t) + (x_{i,j,k} x_{i,j,k}^{c,f}(t) R_{i,j,k}^{c,f}(t) n_{i,j,k}^{c,f}(t)\Big]}{R_k(t-1)} \quad (3.19)$$

subject to (3.12)-(3.15). By gradient-based scheduling, multi-carrier scheduling problem $\mathbf{P}_1$ can be decomposed into multiple single-carrier scheduling problem $\mathbf{P}_2$. $\mathbf{P}_2$ only consists of linear objective function and linear constraints with variables $n_{i,j,k}^{f}(t)$ and $n_{i,j,k}^{c,f}(t)$. Thus it

is a convex problem.

3.4.1 Optimal Resource Scheduling Scheme by Solving the KKT Conditions

For convex optimization problems, the KKT conditions are necessary and sufficient for optimality. Optimal solution for the convex optimization problem $\mathbf{P}_2$ can thus be solved from the KKT conditions. By introducing Lagrangian multipliers $\lambda_i^f(t)$, $\mu_{i,j}^f(t)$, $\nu_{i,j,k}^f(t)$ and $\nu_{i,j,k}^{c,f}(t)$, the Lagrangian function of $\mathbf{P}_2$ is shown in (3.20).

$$\begin{aligned}
&\mathcal{L}\left(n_{i,0,k}^f(t), n_{i,j,k}^f(t), n_{i,j,k}^{c,f}(t), \lambda_i^f(t), \mu_{i,j}^f(t), \nu_{i,j,k}^f(t), \nu_{i,j,k}^{c,f}(t)\right) = \\
-\ &\sum_{k=1}^{N_u}\sum_{i=1}^{N_c}\sum_{j=0}^{N_r}\frac{1}{R_k(t-1)}\Big(x_{i,j,k}(1-x_{i,j,k}^{c,f}(t))R_{i,j,k}^f(t)n_{i,j,k}^f(t) + x_{i,j,k}x_{i,j,k}^{c,f}(t)R_{i,j,k}^{c,f}(t)n_{i,j,k}^{c,f}(t)\Big) \\
+\ &\sum_{i=1}^{N_c}\lambda_i^f(t)\Big(\sum_{k=1}^{N_u}x_{i,0,k}n_{i,0,k}^f(t) + \sum_{j=1}^{N_r}\sum_{k=1}^{N_u}x_{i,j,k}x_{i,j,k}^{c,f}(t)n_{i,j,k}^{c,f}(t) - 1\Big) \\
+\ &\sum_{i=1}^{N_c}\sum_{j=1}^{N_r}\mu_{i,j}^f(t)\Big(\sum_{k=1}^{N_u}x_{i,j,k}(1-x_{i,j,k}^{c,f}(t))n_{i,j,k}^f(t) + \sum_{k=1}^{N_u}x_{i,j,k}x_{i,j,k}^{c,f}(t)n_{i,j,k}^{c,f}(t) - 1\Big) \\
-\ &\sum_{i=1}^{N_c}\sum_{j=0}^{N_r}\sum_{k=1}^{N_u}\nu_{i,j,k}^f(t)n_{i,j,k}^f(t) - \sum_{i=1}^{N_c}\sum_{j=1}^{N_r}\sum_{k=1}^{N_u}\nu_{i,j,k}^{c,f}(t)n_{i,j,k}^{c,f}(t). \qquad (3.20)
\end{aligned}$$

Then the KKT conditions are given as follows:

- **Primal feasibility**

$$\sum_{k=1}^{N_u}x_{i,0,k}n_{i,0,k}^f(t) + \sum_{j=1}^{N_r}\sum_{k=1}^{N_u}x_{i,j,k}x_{i,j,k}^{c,f}(t)n_{i,j,k}^{c,f}(t) \le 1 \qquad (3.21)$$
$$\text{for } f = 1, \cdots, F$$

$$\sum_{k=1}^{N_u}x_{i,j,k}(1-x_{i,j,k}^{c,f}(t))n_{i,j,k}^f(t) + \sum_{k=1}^{N_u}x_{i,j,k}x_{i,j,k}^{c,f}(t)n_{i,j,k}^{c,f}(t) \le 1 \qquad (3.22)$$
$$\text{for } i = 1, \cdots, N_c,\ j = 1, \cdots, N_r, f = 1, \cdots, F$$

$$-n_{i,j,k}^f(t) \le 0 \quad \forall i, j, k, f \qquad (3.23)$$

$$-n_{i,j,k}^{c,f}(t) \le 0 \quad \forall i, j, k, f \qquad (3.24)$$

- **Dual feasibility**

$$\lambda_i^f(t) \geq 0 \quad \forall i, f \tag{3.25}$$

$$\mu_{i,j}^f(t) \geq 0 \quad \forall i, f \tag{3.26}$$

$$\nu_{i,j,k}^f(t) \geq 0 \quad \forall i, j, k, f \tag{3.27}$$

$$\nu_{i,j,k}^{c,f}(t) \geq 0 \quad \forall i, j, k, f \tag{3.28}$$

- **Complementary slackness**

$$\lambda_i^f(t)\Big(\sum_{k=1}^{N_u} x_{i,0,k} n_{i,0,k}^f(t) + \sum_{j=1}^{N_r}\sum_{k=1}^{N_u} x_{i,j,k} x_{i,j,k}^{c,f}(t) n_{i,j,k}^{c,f}(t) - 1\Big) = 0 \tag{3.29}$$
$$\text{for } i = 1, \cdots, N_c, f = 1, \cdots, F$$

$$\mu_{i,j}^f(t)\Big(\sum_{k=1}^{N_u} x_{i,j,k}(1 - x_{i,j,k}^{c,f}(t)) n_{i,j,k}^f(t) + \sum_{k=1}^{N_u} x_{i,j,k} x_{i,j,k}^{c,f}(t) n_{i,j,k}^{c,f}(t) - 1\Big) = 0 \tag{3.30}$$
$$\text{for } i = 1, \cdots, N_c, j = 1, \cdots, N_r, f = 1, \cdots, F$$

$$\nu_{i,j,k}^f(t) n_{i,j,k}^f(t) = 0 \quad \forall i, j, k, f \tag{3.31}$$

$$\nu_{i,j,k}^{c,f}(t) n_{i,j,k}^{c,f}(t) = 0 \quad \forall i, j, k, f \tag{3.32}$$

- **Stationarity**

$$\frac{\partial \mathcal{L}\left(n_{i,0,k}^f(t), n_{i,j,k}^f(t), n_{i,j,k}^{c,f}(t), \lambda_i^f(t), \mu_{i,j}^f(t), \nu_{i,j,k}^f(t), \nu_{i,j,k}^{c,f}(t)\right)}{\partial n_{i,0,k}^f(t)} = 0 \quad \forall i, k, f \tag{3.33}$$

$$\frac{\partial \mathcal{L}\left(n_{i,0,k}^f(t), n_{i,j,k}^f(t), n_{i,j,k}^{c,f}(t), \lambda_i^f(t), \mu_{i,j}^f(t), \nu_{i,j,k}^f(t), \nu_{i,j,k}^{c,f}(t)\right)}{\partial n_{i,j,k}^f(t)} = 0 \quad \forall i, j, k, f \tag{3.34}$$

$$\frac{\partial \mathcal{L}\left(n_{i,0,k}^f(t), n_{i,j,k}^f(t), n_{i,j,k}^{c,f}(t), \lambda_i^f(t), \mu_{i,j}^f(t), \nu_{i,j,k}^f(t), \nu_{i,j,k}^{c,f}(t)\right)}{\partial n_{i,j,k}^{c,f}(t)} = 0 \quad \forall i, j, k, f \tag{3.35}$$

From (3.33)-(3.35), we have

$$-\frac{R^f_{i,0,k}(t)}{R_k(t-1)}x_{i,0,k}+\lambda^f_i(t)x_{i,0,k}-\nu^f_{i,0,k}(t)=0 \tag{3.36}$$

$$\left(-\frac{R^f_{i,j,k}(t)}{R_k(t-1)}+\mu^f_{i,j}(t)\right)x_{i,j,k}\left(1-x^{c,f}_{i,j,k}(t)\right)-\nu^f_{i,j,k}(t)=0 \tag{3.37}$$

$$\left(-\frac{R^{c,f}_{i,j,k}(t)}{R_k(t-1)}+\lambda^f_i(t)+\mu^f_{i,j}(t)\right)x_{i,j,k}x^{c,f}_{i,j,k}(t)-\nu^{c,f}_{i,j,k}(t)=0 \tag{3.38}$$

It is observed that the optimization problem can be decomposed into N_c independent sub-problems, where each corresponds to one cell. Therefore, the resource scheduling problem can be solved for each cell independently and parallelly with multiple threads efficiently. Without loss of generality, we analytically solve the problem for the ith cell, and the proposed solving methodology can be applicable to all the other cells in the system.

Our goal is to find at time t, for each RB f in each sector, the optimal MUE index k_0^*, the optimal PUE index $k^*_{1,j}$, and the optimal CUE index k_2^* for the mBS, the jth pBS, and their joint processing. It also needs to decide the corresponding optimal $n^{f*}_{i,j,k}(t)$ and $n^{c,f*}_{i,j,k}(t)$ values. Towards this end, based on the above KKT conditions [21], we solve the optimal Lagrangian multipliers as follows.

$$\lambda^{f*}_i(t) = \max\{\lambda^f_{i,A}(t),\lambda^f_{i,j*,B}(t)\} \tag{3.39}$$

$$\mu^{f*}_{i,j}(t) = \max_{k_{1,j}\in\mathcal{K}_{i,j}}\frac{R^f_{i,j,k_{1,j}}(t)}{R_{k_{1,j}}(t-1)} \tag{3.40}$$

$$\nu^{f*}_{i,j,k_0}(t) = \lambda^{f*}_i(t)-\frac{R^f_{i,0,k_0}(t)}{R_{k_0}(t-1)} \quad \text{for } k_0\in\mathcal{K}_{i,0} \tag{3.41}$$

$$\nu^{f*}_{i,j,k_{1,j}}(t) = \mu^{f*}_{i,j}(t)-\frac{R^f_{i,j,k_{1,j}}(t)}{R_{k_{1,j}}(t-1)} \quad \text{for } k_{1,j}\in\mathcal{K}_{i,j} \tag{3.42}$$

$$\nu^{c,f*}_{i,j,k_2}(t) = \lambda^{f*}_i(t)-\left[\sum_{k_2\in\mathcal{Q}}\frac{R^{c,f}_{i,j,k_2}(t)}{R_{k_2}(t-1)}-\max_{k_{1,j}\in\mathcal{K}_{i,j}}\frac{R^f_{i,j,k_{1,j}}(t)}{R_{k_{1,j}}(t-1)}\right] \text{for } k_2\in\mathcal{K}^{c,f}_{i,j}(t) \tag{3.43}$$

where

$$\lambda_{i,A}^{f}(t) = \max_{k_0 \in \mathcal{K}_{i,0}} \frac{R_{i,0,k_0}^{f}(t)}{R_{k_0}(t-1)} \tag{3.44}$$

$$\lambda_{i,j*,B}^{f}(t) = \max_{j \in \mathcal{J}} \Big(\max_{\mathcal{Q} \subseteq \mathcal{K}_{i,j}^{c,f}} \sum_{k_2 \in \mathcal{Q}} \frac{R_{i,j,k_2}^{c,f}(t)}{R_{k_2}(t-1)} - \max_{k_{1,j} \in \mathcal{K}_{i,j}} \frac{R_{i,j,k_{1,j}}^{f}(t)}{R_{k_{1,j}}(t-1)} \Big). \tag{3.45}$$

Here, $\mathcal{J}$ is a set of pBSs in each sector, $\mathcal{Q}_{i,j}^{c,f}(t)$ ($\dim \mathcal{Q}_{i,j}^{c,f}(t) = 2$ and $\mathcal{Q}_{i,j}^{c,f}(t) \subseteq \mathcal{K}_{i,j}^{c,f}(t)$) consists of the two CUEs that are jointly processed by the mBS and pBS j on RB f by using the precoding technique.

It can be seen that $\lambda_{i,A}^{f}(t)$ and $\lambda_{i,j*,B}^{f}(t)$ represent the gains in proportional fairness value at the ith mBS by different strategies in assigning the fth RB at time t. Specifically, $\lambda_{i,A}^{f}(t)$ calculates the gain in assigning RB f to the best MUE. $\lambda_{i,j*,B}^{f}(t)$ calculates the gain in assigning RB f to the best CUEs in the coverage range of the pBS j^*. The value of $\lambda_{i}^{f}(t)$ is chosen to be the highest among all the gains under different strategies, and the corresponding UE is assigned with the RB.

Based on the obtained $\lambda_{i}^{f}(t)$ value, the optimal value of $\mu_{i,j}^{f}(t)$ can be calculated from (3.40). It can be considered as the proportional fairness gain at the jth pBS in the ith sector by serving the selected PUE. Specifically, the term $R_{i,j,k_{1,j}}^{f}(t)/R_{k_{1,j}}(t-1)$ is the gain of serving the k_1th PUE by the jth pBS.

Based on the derived optimal Lagrangian multiplier values, we consider the following two cases in finding the optimal k_0^*, $k_{1,j}^*$, k_2^*, $n_{i,j,k}^{f*}(t)$, and $n_{i,j,k}^{c,f*}(t)$ values for each RB f in each sector i at time t .

Case − 1: $\lambda_{i,A}^{f}(t) \geq \lambda_{i,j*,B}^{f}(t)$

In this case, we have

$$\lambda_{i}^{f}(t) = \max_{k_0 \in \mathcal{K}_{i,0}} \frac{R_{i,0,k_0}^{f}(t)}{R_{k_0}(t-1)}, \tag{3.46}$$

and

$$\max_{\mathcal{Q} \subseteq \mathcal{K}_{i,j_1^*}^{c,f}} \sum_{k_2 \in \mathcal{Q}} \frac{R_{i,j_1^*,k_2}^{c,f}(t)}{R_{k_2}(t-1)} < \max_{k_0 \in \mathcal{K}_{i,0}} \frac{R_{i,0,k_0}^{f}(t)}{R_{k_0}(t-1)} + \frac{R_{i,j_1^*,k_{1,j}^*}^{f}(t)}{R_{k_{1,j}^*}(t-1)}, \tag{3.47}$$

where

$$j_1^* = \arg\max_{j\in\mathcal{J}} \Big(\max_{\mathcal{Q}\subseteq\mathcal{K}_{i,j}^{c,f}} \sum_{k_2\in\mathcal{Q}} \frac{R_{i,j,k_2}^{c,f}(t)}{R_{k_2}(t-1)} - \frac{R_{i,j,k_{1,j}^*}^{f}(t)}{R_{k_{1,j}^*}(t-1)} \Big). \tag{3.48}$$

The left hand side of the inequalities (3.47) is the proportional fairness gain by serving the best two CUEs in the fth RB of mBS i. The right hand side of (3.47) is the proportional fairness gain by serving the best MUE on RB f of mBS i and the best PUE associated with the pBS j_1 on RB f separately. From (3.47), we know that the case with $\lambda_{i,A}^f(t) \geq \lambda_{i,j*,B}^f(t)$ corresponds to the scenario where serving the CUE cooperatively on RB f by the mBS and the pBS receives a less gain than using the RB for the MUE and the PUE separately. In another word, CoMP and precoding shall not be used on RB f for mBS i.

Substituting (3.46) into (3.40)-(3.43), we have $\nu_{i,j,k_2}^{c,f*}(t) > 0$ for all $k_2 \in \mathcal{K}_{i,j}^{c,f}(t)$, and we can get the optimal indexes for different UEs.

The optimal MUE index is

$$k_0^* = \arg\max_{k_0\in\mathcal{K}_{i,0}} \frac{R_{i,0,k_0}^{f}(t)}{R_{k_0}(t-1)}, \tag{3.49}$$

and the index of optimal PU is

$$k_{1,j}^* = \arg\max_{k_{1,j}\in\mathcal{K}_{i,j}} \frac{R_{i,j,k_{1,j}}^{f}(t)}{R_{k_{1,j}}(t-1)}. \tag{3.50}$$

In the case with $\lambda_{i,A}^f(t) \geq \lambda_{i,j*,B}^f(t)$, the optimal strategy in allocating the fth RB at the tth subframe in the ith mBS is to let the mBS transmit to the k_0^*th MUE on the entire RB f, the j_1th pBS transmit to the $k_{1,j}^*$th PUE on the entire RB f.

The optimal $n_{i,j,k}^f(t)$ and $n_{i,j,k}^{c,f}(t)$ values for the virtual resource allocation problem can be solved from (3.12) and (3.13) as

$$n_{i,0,k}^{f*} = 0 \text{ for } k \neq k_0^*,\ n_{i,j,k}^{f*} = 0 \text{ for } k \neq k_{1,j}^* \tag{3.51}$$

$$n_{i,j,k}^{c,f*} = 0, \quad n_{i,0,k_0^*}^{f*}(t) = 1, \quad n_{i,j,k_{1,j}^*}^{f*}(t) = 1. \tag{3.52}$$

Case − 2: $\lambda_{i,A}^{f}(t) < \lambda_{i,j*,B}^{f}(t)$

In this case, we have

$$\lambda_i^{f*}(t) = \max_{j\in\mathcal{J}} \Big(\max_{\mathcal{Q}\subseteq\mathcal{K}_{i,j}^{c,f}} \sum_{k_2\in\mathcal{Q}} \frac{R_{i,j,k_2}^{c,f}(t)}{R_{k_2}(t-1)} - \frac{R_{i,j,k_{1,j}^*}^{f}(t)}{R_{k_{1,j}^*}(t-1)} \Big). \tag{3.53}$$

Following the same analysis in Case-1, it is known that $\lambda_{i,A}^{f}(t) < \lambda_{i,j*,B}^{f}(t)$ corresponds to a scenario where the gain in proportional fairness value by serving the CUEs cooperatively on RB f by the mBS and the pBS is higher than the gain in using RB f to serve the MUE and the PUE separately. We can obtain the indexes for optimal CUE

$$k_2^* = \arg \max_{\mathcal{Q}\subseteq\mathcal{K}_{i,j^*}^{c,f}} \sum_{k_2\in\mathcal{Q}} \frac{R_{i,j^*,k_2}^{c,f}(t)}{R_{k_2}(t-1)} \tag{3.54}$$

and

$$j^* = \arg\max_{j\in\mathcal{J}} \Big(\max_{\mathcal{Q}\subseteq\mathcal{K}_{i,j}^{c,f}} \sum_{k_2\in\mathcal{Q}} \frac{R_{i,j,k_2}^{c,f}(t)}{R_{k_2}(t-1)} - \frac{R_{i,j,k_{1,j}^*}^{f}(t)}{R_{k_{1,j}^*}(t-1)} \Big), \tag{3.55}$$

where $k_{j,1}^*$ value is given in (3.52).

In the case with $\lambda_{i,A}^{f}(t) < \lambda_{i,j*,B}^{f}(t)$, the optimal resource allocation strategy at the tth subframe is to allocate the fth RB of the ith mBS and the j^*th pBS to jointly serve the k_2^*th CUE, allocate the fth RB of the pBS with index $j \in \mathcal{J}, j \neq j^*$ to serve the $k_{1,j}^*$th RUE.

The optimal $n_{i,j,k}^{f}(t)$ and $n_{i,j,k}^{c,f}(t)$ values for the virtual resource allocation problem can be solved from (3.12)-(3.13) as

$$n_{i,j,k}^{c,f*} = 0 \text{ for } k \neq k_2^*,\ n_{i,j,k}^{f*} = 0 \text{ for } k \neq k_{1,j}^* \tag{3.56}$$

$$n_{i,0,k}^{f*} = 0, \quad n_{i,j,k_{1,j}^*}^{f*}(t) = 1, \quad n_{i,j,k_2^*}^{c,f*}(t) = 1. \tag{3.57}$$

3.4.2 Summary of Optimization Procedure

So far, we derive the index for optimal MUE, PUE and CUE. For the CUE with the

indexes $k_2 \in \mathcal{K}_{i,j}^{c,f}(t)$, we need to try different combinations to find the best combination of two CUEs $\mathcal{Q}$ to be served simultaneously, which is very complicated and unfeasible in practice. In order to tackle the computational complexity, we propose a three-step optimization procedure.

Step − 1: Determine UE's associations status

1 − 1): Based on the bias value δ, all the UEs in the system can be decided as either MUEs or PUEs.

Step − 2: At time t, form the CoMP candidate set $\mathcal{K}_{i,j}^{c,f}(t)$. The CUEs in the same CoMP set $\mathcal{K}_{i,j}^{c,f}(t)$ should be associated with the same pBS.

2 − 1): Given an SINR threshold α, all the PUE whose SINR values are less than α are marked as CUEs and form the CoMP candidate set.

Step − 3: At time t, apply the proposed resource scheduling in two rounds

3 − 1): In the 1st round, following the discussion in Case-1 and Case-2, we assign the best UE (MUE , PUE or CUE) for each RB at each mBS and pBS.

3 − 2): In the 2nd round, for each RB allocated to CUE at the corresponding mBS and pBS, identify the second CUE in the same CoMP set to share the same RB. Apply (3.4) to obtain the equivalent channel gains. Then evaluate the SINR and R to find the second CUE via (3.54).

3.5 Performance Evaluation

We conduct the simulation study in a 19-cell 3-sector three-ring hexagonal cellular network with the extended typical urban (ETU) channel model. Simulation setup follows the guidelines described in 3GPP technical report [22]. The total bandwidth is 10MHz and 180 kHz for each resource block (RB). There are 50 RBs in each frequency band. The transmit power of the mBS is 46dBm (40W) and the transmit power of the pBS is 30dBm (1W). 50 UEs are uniformly distributed in each sector and travel at a speed of 3 km/h.

In Fig. 3.2, we evaluate the performance of the systems with and without CoMP, as well as with and without THP. We express the system throughput as the relative percentage

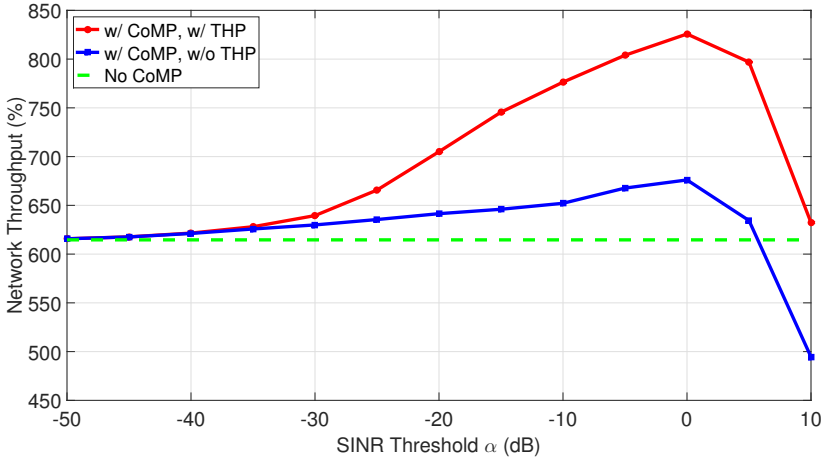

Fig. 3.2: Network throughput comparison at bias value $\delta = 0$ dB

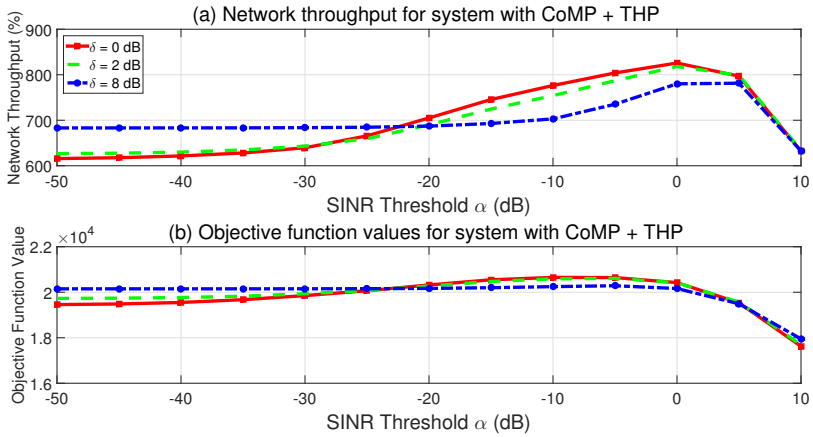

Fig. 3.3: Performance comparison of system with CoMP and THP at different δ

of the throughput of the homogeneous network, which consists of only one mBS per sector. Worth of mention, the system throughput does not change with SINR threshold α if no CoMP is applied. With CoMP, the system throughput reaches the maximum when $\alpha^* =$ 0dB. The throughput goes lower when α is either higher or lower than 0dB. When α is lower, poor SINR UEs may not receive CoMP and their low throughput lead to overall lower system throughput. When α is higher, more UEs in the pBS cell will become CUEs. For CUEs with poor SINRs, they have a very good chance to be served cooperatively by mBS and pBS. However, if α is too high, UEs with good SINRs are unnecessarily served by mBS and pBS cooperatively, leading to a waste of radio resources. Hence, we can see

the throughput falloff when α exceeds 0dB. We also compare the CoMP schemes with and without THP. With THP, mBS and pBS can support two CUEs simultaneously on the same RB. The system with CoMP + THP achieves a much larger capacity gain than with CoMP only, approximately 830% vs 680% when $\alpha^* = 0$dB.

In Fig. 3.3, we compare the system throughput performances with CoMP + THP under different mobile association biases (δ values). $\delta = 0$dB corresponds to a pathloss-based mobile association while $\delta = 16$dB represents a best-power based mobile association. With different δ values, the system can always achieve the highest throughput gain at $\alpha^* = 0$dB. If we evaluate the system in terms of log scale throughput, which is the objective function of $\mathbf{P}_2$, the system log scale throughput is maximized at $\alpha^* = -5$dB. The system can benefit more from a high α threshold but also suffers more from a low α when δ value is low. Comparing $\delta = 2$dB and $\delta = 8$dB, more UEs will fall into the coverage area of pBS when $\delta = 2$dB and it will lead to a larger pool of CUEs participating CoMP. A too low α will leave a large number of UEs at the pBS cell edge at low SINR region, thus leading to a low system throughput. This particularly hurts the scenario using a low mobile association bias δ. Therefore, SINR threshold α needs to be selected appropriately in order to realize a high capacity gain.

3.6 Chapter Summary

In this chapter, a precoding-based CoMP transmission scheme is proposed to optimize radio resource scheduling in OFDMA-based heterogeneous networks. The proposed scheme applied Tomlinson-Harashima precoding technique to cancel out the inter-user interference, and combines it with CoMP technique so that the network capacity and cell edge user experience can be improved considerably. Extensive simulations are conducted to investigate the impacts of association schemes on the system performance, and show the performance gains achieved by the proposed scheme.

Chapter 4

Hybrid MU-MIMO and Non-orthogonal Multiple Access Design in Wireless Heterogeneous Networks

4.1 Introduction

In the previous chapters, we mainly aim to avoid or mitigate intra-cell and inter-cell interferences in order to improve cell edge user experience and spectrum efficiency. Specifically, we sketch different practical coordination schemes, with advanced precoding approach, to assess their performance in system-level simulations. In this chapter, we continue to explore the potential of network coordination scheme by incorporating with other promising techniques.

As a future radio access scheme — Non-orthogonal Multiple Access (NOMA) is first proposed by DoCoMo for 5G networks [30]. It is based on conventional Orthogonal Frequency-Division Multiple Access (OFDMA) or discrete Fourier transform (DFT)-spread OFDM. The fundamental idea behind NOMA is to explore the power domain for multiple access [31]. Instead of using orthogonal spectrum, NOMA allocates the same spectrum to different users, where different users are served with different power levels. Therefore, the access scheme can potentially achieve a high spectrum efficiency and increase the total system throughput considerably. NOMA enables multiple users to share the same spectrum resource simultaneously by doing interference cancellation on the receiving side. Thus advanced multiuser detection and interference cancellation techniques are required to retrieve the signals at the receivers. In [32], successive interference cancellation (SIC) is used to extract the intended signal from the received aggregated data. The system-level study demonstrated the performance gain of NOMA over traditional orthogonal multiple access techniques.

In this chapter, we propose a new framework that considers a hybrid design of NOMA

and MU-MIMO. It is shown in [31] [32] that there exist some limitations in NOMA. For UEs with a relative large power disparity, NOMA is able to achieve a good performance gain over non-NOMA case. When the power difference becomes small between two received signals, NOMA gain diminishes. Alternatively, MU-MIMO can work well under this situation given that there is enough channel diversity. More specifically, we introduce a hybrid MU-MIMO and NOMA design scheme in wireless heterogeneous networks to improve the system throughput and also to increase multi-user diversity gains by exploiting the heterogeneous nature of the supporting wireless networks. The best user cluster is formed in a NOMA group and then a precoding-based MU-MIMO scheme is applied to NOMA composite signals. The problem is further formulated as a resource scheduling optimization problem with proportional fairness purpose. Aiming to ensure the global optimality, a brute-force search algorithm is used to solve the problem.

4.2 Hybrid MU-MIMO and NOMA Framework

We consider downlink communications in a wireless heterogeneous network in Fig. 4.1. As a preliminary study, we assume all the UEs and BSs are equipped with single antenna. Frequency reuse one is deployed throughout the system. The overlaid pico-cells reuse the same spectrum of the macro-cells and aim to provide services locally with less energy, mainly at hotspots and coverage holes, such that the overall system spectrum efficiency, energy efficiency and coverage are greatly improved.

As mentioned in [31] [32], NOMA usually achieves great performance gains if there exist relatively large disparities between the received signals among a cluster of users. For users with little difference in received signal strengths, NOMA might not provide any performance gain. As illustrated in Fig. 4.1, we categories overall service areas into three ranges: mBS MU-MIMO+NOMA range, mBS+pBS MU-MIMO range, and pBS MU-MIMO+NOMA range, based on the association schemes and the received powers from mBS and pBS. For UEs located in the MU-MIMO range, UEs receive relatively equal signal powers from both mBS and pBS and thus MU-MIMO is favorable. For UEs located in MU-MIMO+NOMA ranges, signals received from mBS and pBS are largely different, mak-

ing hybrid MU-MIMO+NOMA as a more spectrum efficient transmission mechanism than either NOMA alone or MU-MIMO alone. In a wireless heterogeneous network, due to the high disparity on the powers from different BSs, a high percentage of downlink UEs locate in the regions where interference power is even stronger than the intended signal power. So using a hybrid MU-MIMO+NOMA scheme in these regions is highly desirable since it turns a destructive interference issue into a constructive contributor. In this paper, for the sake of clarity on presentation, we assume all the UEs and BSs have only one antenna. The algorithm is applied to the general multi-antenna case as well.

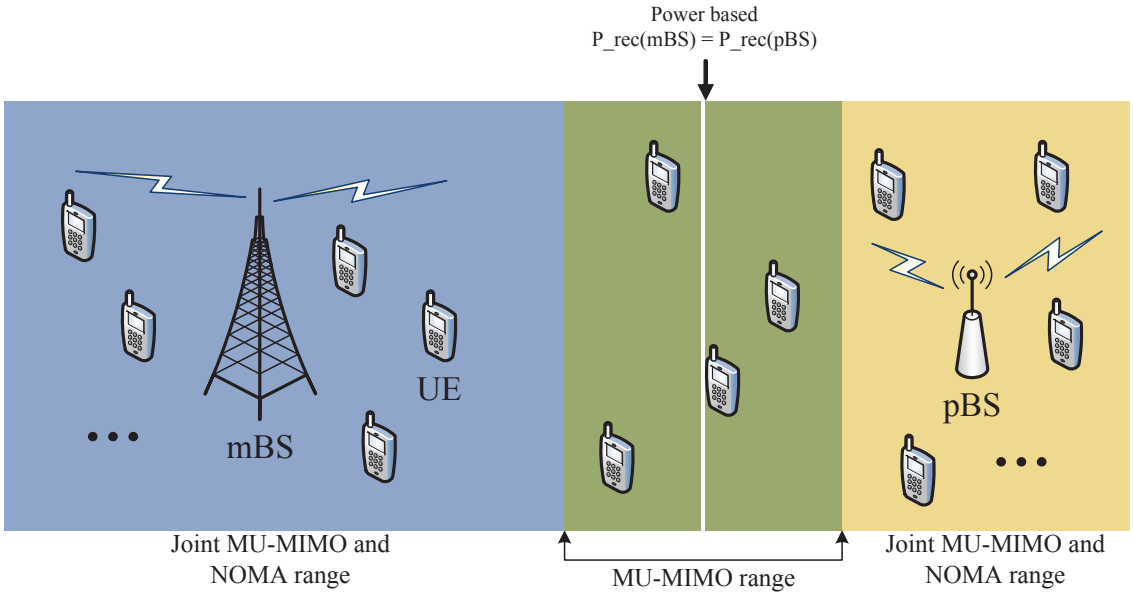

Fig. 4.1: Wireless network model

4.2.1 MU-MIMO

We first address how MU-MIMO works in the MU-MIMO only range. Without loss of generality, we formulate the channel matrix for two users that form an MU-MIMO pair as $\mathbf{H}_{1,2}$ as:

$$\mathbf{H}_{1,2} = \begin{bmatrix} h_{1,1} & h_{1,2} \\ h_{2,1} & h_{2,2} \end{bmatrix}, \tag{4.1}$$

where channel gain $h_{i,j}$ considers both large scale path-loss and small scale Rayleigh fading.

We can simply use $\mathbf{H}$ to represent $\mathbf{H}_{i,j}$. The received downlink signal vector at two UEs can be expressed as:

$$\mathbf{y} = \mathbf{H}\mathbf{x} + \mathbf{z}, \tag{4.2}$$

where $\mathbf{z} = [z_1 \quad z_2]^T$ represents the noise vector on the receiving side. In order to cancel out inter-user interference, we assume perfect channel state information (CSI) is known at both mBS and pBS. Then we apply dirty paper coding (DPC) [33] by designing the precoding matrix as:

$$\mathbf{W} = \mathbf{Q}^H\mathbf{G}, \tag{4.3}$$

where $\mathbf{Q}^H$ is the Hermitian matrix of $\mathbf{Q}$, which is obtained by proceeding LQ decomposition to $\mathbf{H}$:

$$\mathbf{H} = \begin{bmatrix} l_{1,1} & 0 \\ l_{2,1} & l_{2,2} \end{bmatrix} \begin{bmatrix} \mathbf{q_1} \\ \mathbf{q_2} \end{bmatrix} = \mathbf{L}\mathbf{Q}. \tag{4.4}$$

Here, $\mathbf{L}$ is a lower triangle matrix. In order to form an interference free transmission channel, $\mathbf{G}$ is given as

$$\mathbf{G} = \begin{bmatrix} 1 & 0 \\ -\frac{l_{2,1}}{l_{2,2}} & 1 \end{bmatrix}. \tag{4.5}$$

At mBS and pBS, the signal vector $\mathbf{x} = [x_1 \quad x_2]^T$ is precoded to $\hat{\mathbf{x}} = \mathbf{W}\mathbf{x} = [\hat{x}_1 \quad \hat{x}_2]^T$ before transmission, where $\hat{x}_1$ is the precode signal transmitted from mBS and $\hat{x}_2$ is from pBS. Thereby, the received signal $\mathbf{y} = [y_1 \quad y_2]^T$ is expressed as

$$\mathbf{y} = \mathbf{H}\hat{\mathbf{x}} + \mathbf{z} = \mathbf{H}\mathbf{W}\mathbf{x} + \mathbf{z} = \begin{bmatrix} l_{1,1} & 0 \\ 0 & l_{2,2} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \begin{bmatrix} z_1 \\ z_2 \end{bmatrix}. \tag{4.6}$$

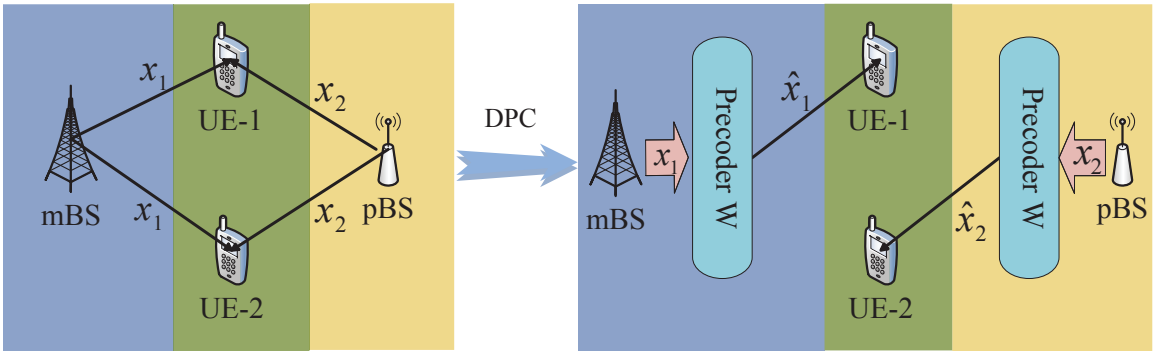

Fig. 4.2: Transmission model for MU-MIMO only

It is observed that the inter-user interference can be perfectly canceled out, which is illustrated in Fig. 4.2. Then the total achievable data rate from MU-MIMO is R^c.

$$R^c = R^c_{k_1} + R^c_{k_2}, \tag{4.7}$$

$$R^c_{k_1} = W \log_2 \left(1 + \frac{|l_{1,1}|^2 P_m}{N_0}\right), \tag{4.8}$$

$$R^c_{k_2} = W \log_2 \left(1 + \frac{|l_{2,2}|^2 P_p}{N_0}\right). \tag{4.9}$$

Here, $l_{m,m}$, $m = 1, 2$, represents the equivalent channel gain between UE 1 (or UE 2) and mBS (or pBS). W is denoted as the bandwidth of one RB.

4.2.2 Hybrid MU-MIMO and NOMA

In the MU-MIMO+NOMA range, we consider an MU-MIMO+NOMA pair which consists of 2 UEs, one from mBS blue range and one from pBS yellow range, shown in Fig. 4.3. Owing to hybrid MU-MIMO and NOMA, each mBS can transmit two signals x_1 and x_2, one for UE 1 and one for UE 2. Each pBS also transmits two signals x_3 and x_4, one for UE 1 and one for UE 2. From each BS's perspective, it transmits two signals to two UEs, one close to itself and one far away, naturally forming a desirable NOMA pair. With hybrid MU-MIMO and NOMA, two signals x_1 and x_4 are intended to UE 1, and the intended signals for UE 2 are x_2 and x_3. So in total 4 signals are transmitted to two UEs by using hybrid MU-MIMO+NOMA, compared with only 2 signals to 2 UEs in the MU-MIMO only

case. In order for the BS to transmit two different signals from a single antenna, power disparity between transmitted signals needs to be imposed in order for the receiving side to achieve a notable NOMA gain. Therefore, we introduce a power allocation parameter $\theta \in (0, 1)$, which is used to partition the transmit power among NOMA signals at each BS.

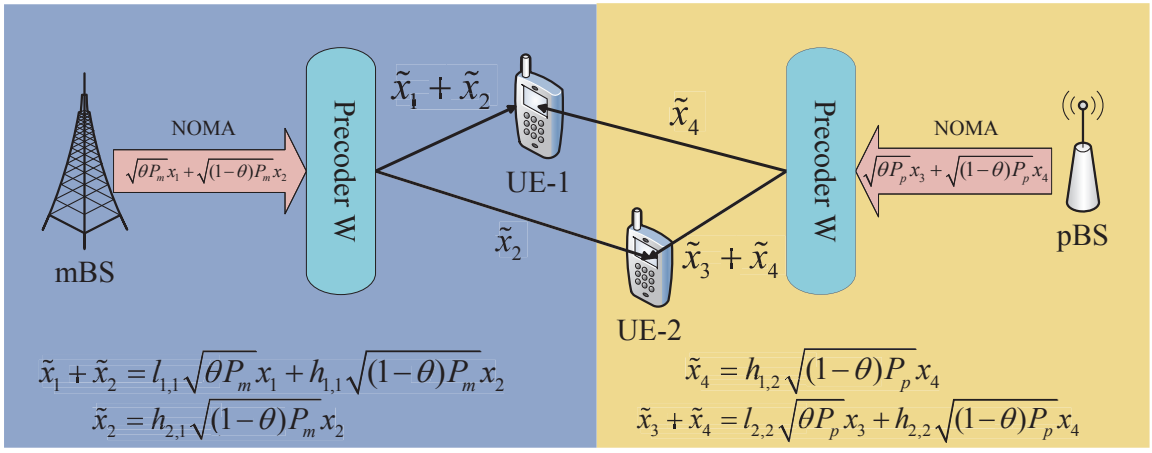

Fig. 4.3: Transmission model for hybrid MU-MIMO and NOMA

On the transmitting side, each BS first uses NOMA to superimpose two signals together and further applies a precoding algorithm to the superimposed signal before sending it out. Without precoding, UE 1 will receive x_1 and x_4 as intended signals while receive x_2 and x_3 as interference. With precoding, UE 1 will still receive x_1 and x_4 as intended signals. But its interference signal reduces to x_3 only. The same applies to UE 2 as well.

As shown in Fig. 4.3, by transmitting the precoded signal, the received signal vector $\mathbf{y}$ is expressed as

$$\begin{aligned}
\mathbf{y} &= \mathbf{H}\hat{\mathbf{x}}_{1,3} + \mathbf{H}\mathbf{x}_{2,4} + \mathbf{z} = \mathbf{H}\mathbf{W}\mathbf{x}_{1,3} + \mathbf{H}\mathbf{x}_{2,4} + \mathbf{z} = \mathbf{L}\mathbf{Q}\mathbf{Q}^H\mathbf{G}\mathbf{x}_{1,3} + \mathbf{H}\mathbf{x}_{2,4} + \mathbf{z} \\
&= \begin{bmatrix} l_{1,1} & 0 \\ 0 & l_{2,2} \end{bmatrix} \begin{bmatrix} \sqrt{\theta P_m}x_1 \\ \sqrt{\theta P_p}x_3 \end{bmatrix} + \begin{bmatrix} h_{1,1} & h_{1,2} \\ h_{2,1} & h_{2,2} \end{bmatrix} \begin{bmatrix} \sqrt{(1-\theta)P_m}x_2 \\ \sqrt{(1-\theta)P_p}x_4 \end{bmatrix} + \begin{bmatrix} z_1 \\ z_2 \end{bmatrix} \\
&= \begin{bmatrix} l_{1,1}\sqrt{\theta P_m}x_1 + h_{1,1}\sqrt{(1-\theta)P_m}x_2 + h_{1,2}\sqrt{(1-\theta)P_p}x_4 \\ l_{2,2}\sqrt{\theta P_p}x_3 + h_{2,1}\sqrt{(1-\theta)P_m}x_2 + h_{2,2}\sqrt{(1-\theta)P_p}x_4 \end{bmatrix} + \begin{bmatrix} z_1 \\ z_2 \end{bmatrix}. \qquad (4.10)
\end{aligned}$$

Here, we assume the first row is the received signal at UE 1 in an MU-MIMO+NOMA pair, and the second row is the received signal at UE 2 in an MU-MIMO+NOMA pair. DPC is used to reduce MU-MIMO inter-user interference. By using DPC, one interfering signal is canceled out at each UE. Thus each UE receives a composite signal consisting of three signals, two of which are intended signals. Further by applying SIC at the receiving side, we can decode three signals sequentially by following the decreasing order of the received signal strength. For the sake of presentation clarity, we denote A as the sum power of the received signals that have power lower than x_1, B as the sum power of the received signals that have power lower than x_2, C as the sum power of the received signals that have power lower than x_3, and D as the sum power of the received signals that have power lower than x_4. Then the total achievable data rate R^n using the hybrid MU-MIMO and NOMA can be expressed as

$$R^n = R^n_{k_1} + R^n_{k_2}, \tag{4.11}$$

$$R^n_{k_1} = W \log_2 \left(1 + \frac{l^2_{1,1}\theta P_m}{A + N_o}\right) + W \log_2 \left(1 + \frac{h^2_{1,2}(1-\theta)P_p}{D + N_o}\right), \tag{4.12}$$

$$R^n_{k_2} = W \log_2 \left(1 + \frac{h^2_{2,1}(1-\theta)P_m}{B + N_o}\right) + W \log_2 \left(1 + \frac{l^2_{2,2}\theta P_p}{C + N_o}\right). \tag{4.13}$$

4.3 Problem Formulation

We intend to design a dynamic transmission mechanism that can maximize the overall system throughput and deliver satisfactory user experience. Towards that end, at each scheduling cycle, the scheme needs to:

- properly decide MU-MIMO with/without NOMA group pair;
- properly adjust the transmit power allocation factor θ to maximize MU-MIMO+NOMA performance gain;
- properly allocate RBs to UE pairs.

The algorithm will need to select the transmission mode, i.e., either MU-MIMO only or hybrid MU-MIMO and NOMA for each RB of each BS among all the candidate UEs in the system. Once the BS pair (mBS i, pBS (i,j)) is determined, the rest pBSs (i,j'), $\forall j' \neq j$, will switch to muting mode so that there is no intra-cell interference. Here mBS i represents the mBS in cell i and pBS (i,j) represents the jth pBS in cell i. Each cell can have multiple pBSs. Thus the following scheduling variables are defined.

$$x^c_{i,0,k_1}(f,t) = \begin{cases} 1; & \text{UE } k_1 \text{ is served by mBS } i \text{ on } f \text{ as the 1st UE in an MU-MIMO} \\ & \text{pair at } t, \\ 0; & \text{otherwise.} \end{cases}$$

$$x^c_{i,j,k_2}(f,t) = \begin{cases} 1; & \text{UE } k_2 \text{ is served by pBS } (i,j) \text{ on } f \text{ as the 2nd UE in an MU-MIMO} \\ & \text{pair at } t, \\ 0; & \text{otherwise.} \end{cases}$$

$$x^n_{i,0,k_1}(f,t) = \begin{cases} 1; & \text{UE } k_1 \text{ is served by mBS } i \text{ on } f \text{ as the 1st UE in a hybrid} \\ & \text{MU-MIMO and NOMA pair at } t, \\ 0; & \text{otherwise.} \end{cases}$$

$$x^n_{i,j,k_2}(f,t) = \begin{cases} 1; & \text{UE } k_2 \text{ is served by pBS } (i,j) \text{ on } f \text{ as the 2nd UE in a hybrid} \\ & \text{MU-MIMO and NOMA pair at } t, \\ 0; & \text{otherwise.} \end{cases}$$

Furthermore, in order to determine whether a RB should be assigned to an MU-MIMO pair or a hybrid MU-MIMO and NOMA pair, we introduce the following proportional fairness (PF) function:

$$U_k(t) = \frac{R_k^\alpha(t)}{T_k^\beta(t)}, \tag{4.14}$$

where

$$T_k(t) = \frac{1}{T_c} \sum_{\tau=t-T_c+1}^{t} R_k(\tau). \tag{4.15}$$

Here, $R_k(t)$ is denoted as the instantaneous data rate of UE pair k_1 and k_2 at time t, and $T_k(t)$ is denoted as the window-based moving average throughput for UE pair at time t. T_c is the moving average window size. α and β tune the "fairness" of the scheduler. From (4.14), if a UE gets a low throughout in the past, its PF function value $U_k(t)$ will be elevated so that its priority to be served increases. Thereby, by properly adjusting α and β, we can ensure that when maximizing spectrum efficiency, UEs in different regions can still be served fairly.

According to (4.7) and (4.11), the achievable data rate at time t on RB f is expressed as:

$$\begin{aligned} R_k(t) &= x^c_{i,0,k_1}(f,t)(1-x^n_{i,0,k_1}(f,t))R^c_{k_1} + x^c_{i,j,k_2}(f,t)(1-x^n_{i,j,k_2}(f,t))R^c_{k_2} \\ &+ x^n_{i,0,k_1}(f,t)(1-x^c_{i,0,k_1}(f,t))R^n_{k_1} + x^n_{i,j,k_2}(f,t)(1-x^c_{i,j,k_2}(f,t))R^n_{k_2}. \end{aligned} \tag{4.16}$$

The objective function of the scheduling problem is thus formulated as

$$[\mathbf{P}_1] \ \max_{\mathbf{x}(t)} \sum_k U_k(t) \tag{4.17}$$

subject to

$$\sum_{k=1}^{N_u} x^c_{i,0,k_1}(f,t) + x^n_{i,0,k_1}(f,t) \leq 1, \ \forall i, f \tag{4.18}$$

$$\sum_{k=1}^{N_u} x^c_{i,j,k_2}(f,t) + x^n_{i,j,k_2}(f,t) \leq 1, \ \forall i, j, f \tag{4.19}$$

Constraints (4.18) and (4.19) ensure that at each time slot, each RB can be assigned to only one pair of UEs, either an MU-MIMO+NOMA pair or an MU-MIMO pair.

Algorithm 1 Brute-force Search Algorithm

Initialization: Given total number of UEs N_u, generate all possible UE pairs. Denote the set of total pairs as $\mathcal{P}_{N_u}$
Convergence = false.
for $t = t_0$ to T **do**
  for $f = 1$ to F **do**
    for $p = 1$ to $|\mathcal{P}_{N_u}|$ **do**
      Identify UE pair indexes $(k_1, k_2)_p \in \mathcal{P}_{N_u}$
      Assume a MU-MIMO pair
      Calculate the objective function value:

$$U_p^c = U_{k_1} + U_{k_2} \tag{4.20}$$

      if $U_p^c \geq U_{p-1}^c$ **then**
        $U_{p^*}^c = U_p^c$
        $(k_1^c, k_2^c) = (k_1, k_2)_p$
      else
        $U_{p^*}^c = U_{p-1}^c$
        $(k_1^c, k_2^c) = (k_1, k_2)_{p-1}$
      end if
      Assume a MU-MIMO+NOMA pair
      Calculate the objective function value:

$$U_p^n = U_{k_1} + U_{k_2} \tag{4.21}$$

      if $U_p^n \geq U_{p-1}^n$ **then**
        $U_{p^*}^n = U_p^n$
        $(k_1^n, k_2^n) = (k_1, k_2)_p$
      else
        $U_{p^*}^n = U_{p-1}^n$
        $(k_1^n, k_2^n) = (k_1, k_2)_{p-1}$
      end if
    end for
    Compare $U_{p^*}^c$ and $U_{p^*}^n$
    Determine the transmission mode:

$$(k_1^{t,f}, k_2^{t,f}) = \arg_{k_1,k_2} \left\{ U_{p^*}^c(k_1^c, k_2^c), U_{p^*}^n(k_1^n, k_2^n) \right\} \tag{4.22}$$

    Assign RB f to UE pair $(k_1^{t,f}, k_2^{t,f})$
    Update average data rate $T_k(t)$
  end for
  Update average data rate $T_k(t)$
end for
Output UEs' average data rates, UEs' transmission modes

4.4 Brute-force Search Algorithm

As a preliminary study on the hybrid MU-MIMO and NOMA framework, we apply a brute-force search algorithm to solve the aforementioned problem. Specifically, in each resource scheduling circle, we search all the UE pairs and form them as either MU-MIMO pairs or MU-MIMO+NOMA pairs. Then we compute their objective function values, and choose the UE pair with highest value as the solution. For additional clarity, we summarize the brute-force search algorithm in Algorithm 1.

4.5 Performance Evaluation

The simulation was set up based on 3GPP case 1 configurations specified in [22]. A single cell network structure is divided into three sectors by 120 degree equally. Each sector represents a macro-cell, in which one mBS is located in the center and 4 pBSs are equally-distanced deployed in the overlaid pico-cells within each macro-cell, forming a two-tier heterogeneous network. UEs are uniformly distributed in the network. Small scale fading is generated based on the Rayleigh fading channel model [34]. Other parameter settings are shown in Table 4.1.

In Fig. 4.4, we investigate the MU-MIMO+NOMA performance at different power allocation factors. It is observed that as θ decreases, UEs have relatively higher average data rates. For example, compared to $\theta = 0.6$, about 80% of the total UEs at $\theta = 0.2$ have an increase of 5000 kbps in average data rates. This is because a small θ reflects a relatively large power disparity within a UE pair. Thereby, implementation of hybrid MU-MIMO and NOMA can deliver additional information to UEs and improve the system throughput considerably. In contrast, with the increase of θ, the received power disparity is not distinct. Then MU-MIMO+NOMA no longer contributes notable performance gains. Hence, more UEs are formed as MU-MIMO pairs. When $\theta = 1$, the system evolves into a pure MU-MIMO system.

Specifically, Fig. 4.5 compares the performance of MU-MIMO users with the hybrid MU-MIMO+NOMA users. It is shown that hybrid MU-MIMO+NOMA users have relatively higher average data rates than MU-MIMO users. This is because the existence

Table 4.1: Simulation parameter settings

Parameter	**Settings**
mBS	1
pBS	4 per macro-cell
UE	200 per cell
Transmitting Antenna	1 per BS
Receiving Antenna	1 per UE
Transmit Power	$P_m = 30$ Watt, $P_p = 1$ Watt
System Bandwidth	10 MHz
Number of RBs	$F = 50$
Bandwidth of RB	$W = 180$ kHz
Size of Time Window	$T_c = 100$ seconds
Fast Fading Model	Rayleigh Fading Channel [34]
Path loss from mBS to UE	$LOS(d) = 103.4 + 24.2 \log_{10}(d)$ $NLOS(d) = 131.1 + 42.8 \log_{10}(d)$
Path loss from pBS to UE	$LOS(d) = 103.8 + 20.9 \log_{10}(d)$ $NLOS(d) = 145.4 + 37.5 \log_{10}(d)$
Shadowing	8 dB, log-normal std. deviation
Noise Model and Density	AWGN, -174 dBm/Hz

of NOMA introduces the diversity gains, and additional information can be transmitted to UEs with the sharing spectrum resources. Therefore, it leads to a leap on the system performance in terms of UEs' data rates.

Fig. 4.6 depicts the system performance at different PF parameters. From (4.14), a large α means that the system is less concerned about user fairness, but more concerned about users with good channel conditions. It can be observed that when $\alpha = 2$ and $\beta = 1$, there exists a relatively large gap between UEs with high data rates and ones with low data rates. This is due to the lack of proportional fairness. UEs with good channel conditions are likely to be served frequently. Conversely, when β is large, the system is more concerned

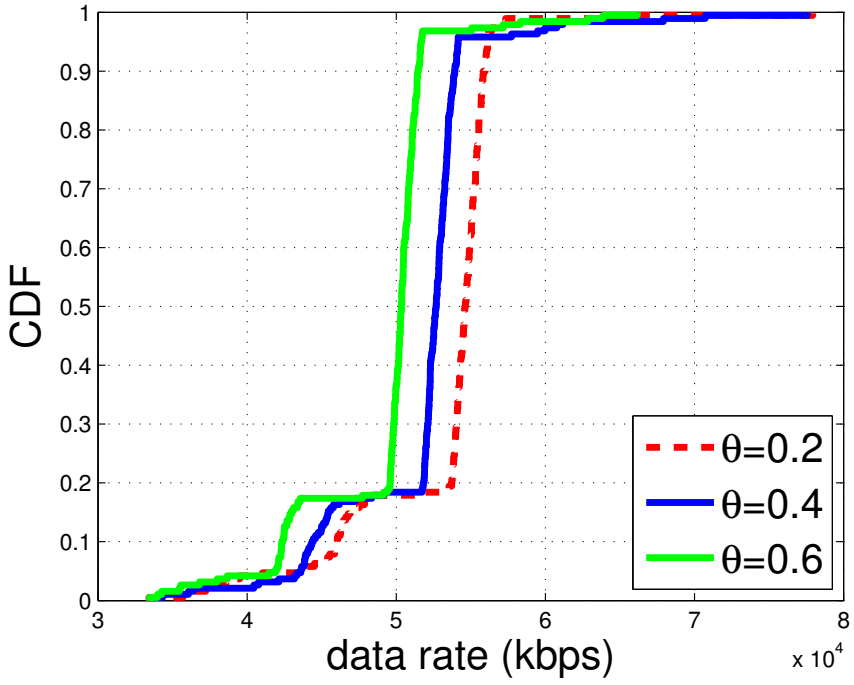

Fig. 4.4: The CDF of user average data rate at different power allocation factor θ

about the average rates. As a result, most of the UEs have the similar average data rates, which is illustrated by the green curve of $\alpha = 1, \beta = 2$.

4.6 Chapter Summary

In this chapter, we investigate a hybrid MU-MIMO and NOMA scheme in a wireless heterogeneous network. A proportional fair resource scheduling problem is formulated to justify the advantage of hybrid scheme. A brute-force search algorithm is applied to solve the problem. Simulation results show that the heterogeneous network can receive considerable benefits from the hybrid MU-MIMO and NOMA implementation. In the future, brute-force search might be inadequate due to its high computational complexity and time consumption. Therefore, for the large-scale system design problem, it is necessary to explore advanced scheduling and pairing methods. Moreover, it is also necessary to consider the hybrid application of MU-MIMO and NOMA in multi-antenna multi-cell systems with inter-cell interferences.

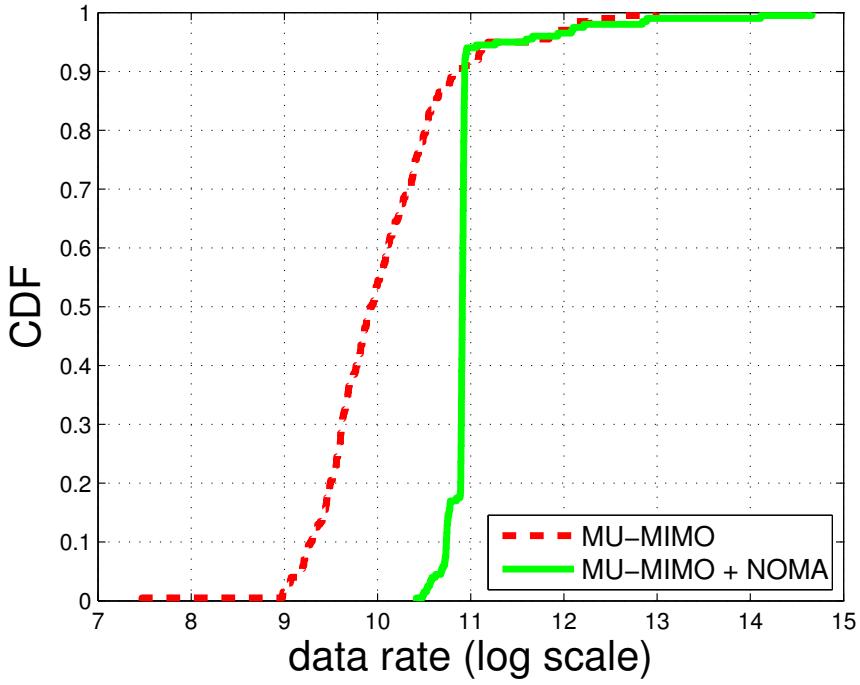

Fig. 4.5: Performance comparison between MU-MIMO and MU-MIMO + NOMA

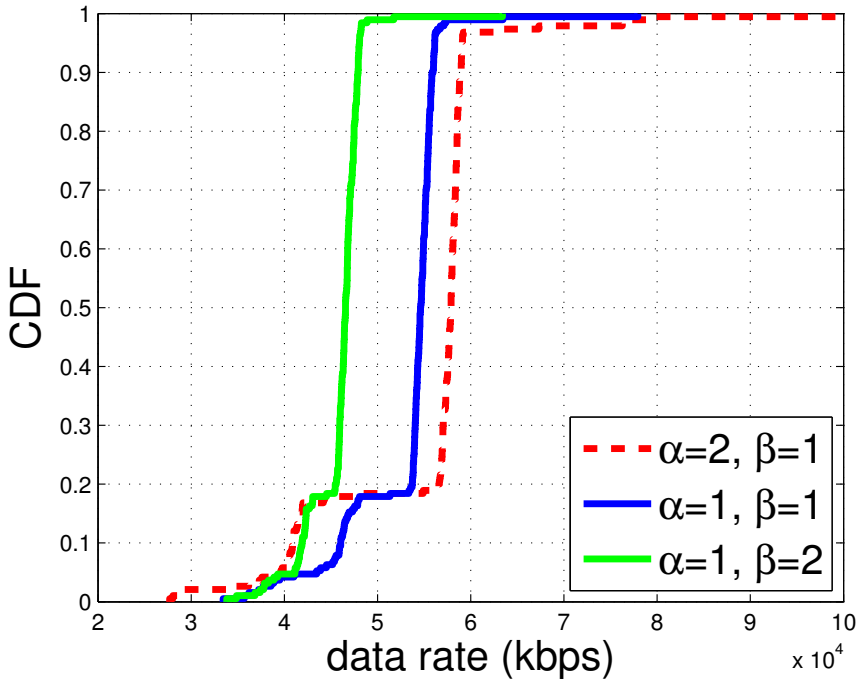

Fig. 4.6: Performance comparison at different PF parameters

Chapter 5

Cooperative Non-orthogonal Multiple Access in Heterogeneous Networks

5.1 Introduction

In this chapter, we consider a multi-antenna system and propose a cooperative NOMA scheme in a wireless heterogeneous network by exploiting the distinct power disparity between mBSs and pBSs. NOMA allows a single UE to receive different data from an mBS and a pBS simultaneously. Furthermore, each BS is equipped with multiple transmitting antennas so that they can cooperatively serve multiple UEs by using DPC to minimize inter-user interference. The proposed resource scheduling policy form UE cooperative cluster properly to maximize the cooperative gain. On the receiving side, SIC is applied to retrieve signals for each transmitting BS sequentially. The resource scheduling optimization problem is formulated as a combinatorial problem. To reduce computational complexity, we introduce a genetic algorithm to form the cooperative cluster and determine serving BSs at each scheduling circle. The performance of the cooperative NOMA-SIC scheme is evaluated and compared with the performance of NOMA-SIC only scheme and with the performance of DPC cooperation only scheme. To the best of our knowledge, there is no similar work in this realm by jointly considering cooperation transmission and NOMA among different BSs.

5.2 Cooperative NOMA Network Model

Without loss generality, we consider a downlink data transmission in a wireless heterogeneous network, as shown in Fig. 5.1. Each cell hosts one mBS and several overlaid pBSs. Each mBS or pBS is equipped with M transmitting antennas and each UE is equipped with one receiving antenna. The total frequency band is divided into F RBs and the size of a

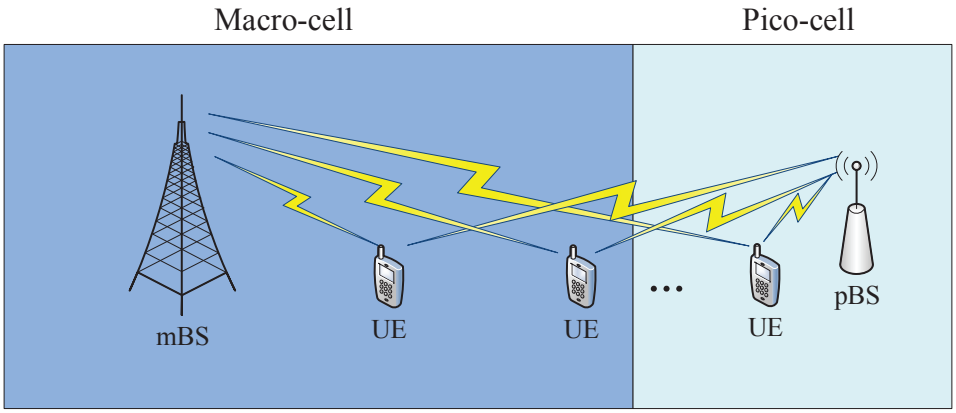

Fig. 5.1: Cooperative NOMA network model

RB represents the scheduling granularity. Therefore, due to DPC technique, there can be up to M UEs served by each BS at each RB. Furthermore, NOMA can allow each UE to receive different information from two BSs including one mBS and one pBS simultaneously. Thus we define the cooperative set $\mathcal{K}_{i,j}^{f}(t)$ ($|\mathcal{K}_{i,j}^{f}(t)| = M$) that consists of UEs cooperatively served by mBS i and pBS (i,j) on RB f at t. pBS (i,j) represents the jth pBS inside cell i.

Let $\mathbf{x}_i$ denote an $M \times 1$ signal vector transmitted from M antennas of mBS i, and $\mathbf{x}_j$ denote an $M \times 1$ signal vector transmitted from M antennas from pBS (i,j). Moreover, $\mathbf{H}_i$ denotes an $M \times M$ channel matrix between mBS i and M receiving UEs, and $\mathbf{H}_j$ denotes an $M \times M$ channel matrix between pBS (i,j) and M receiving UEs. Thus, if mBS i and pBS (i,j) both transmit to the UE set $\mathcal{K}_{i,j}^{f}(t)$ on the same frequency, the received signal can be expressed as:

$$\mathbf{y}_{i,j,\mathcal{K}_{i,j}^{f}(t)} = \mathbf{H}_i\mathbf{x}_i + \mathbf{H}_j\mathbf{x}_j + \mathbf{z}. \tag{5.1}$$

Here, $\mathbf{z}$ is an $M \times 1$ noise vector at receiving UEs. $\mathbf{y}_{i,j,\mathcal{K}_{i,j}^{f}(t)}$ is an $M \times 1$ vector in which each element y_{i,j,m_k}, $1 \leq m_k \leq M$, represents the composite received signal at UE k. k is UE's global index while m_k is the local sequence index for UE k within the cooperative set $\mathcal{K}_{i,j}^{f}(t)$.

5.3 Cooperative NOMA Scheme

By exploiting the distinct power disparity between mBSs and pBSs, a UE with a single antenna can receive simultaneously from an mBS and a pBS with each transmitting a different signal. Furthermore, each BS is equipped with multiple transmitting antennas so that it can serve multiple UEs by using DPC to minimize inter-user interference. SIC scheme is proceeded on receiving side to decode signals from different BSs in a sequential manner. As a summary, DPC is applied for transmissions from a BS with multiple antennas to different users while NOMA is applied to transmissions from different BSs to a single UE. The proposed cooperative NOMA scheme uses both DPC and NOMA to enhance the system throughput, which can be illustrated in Fig. 5.2.

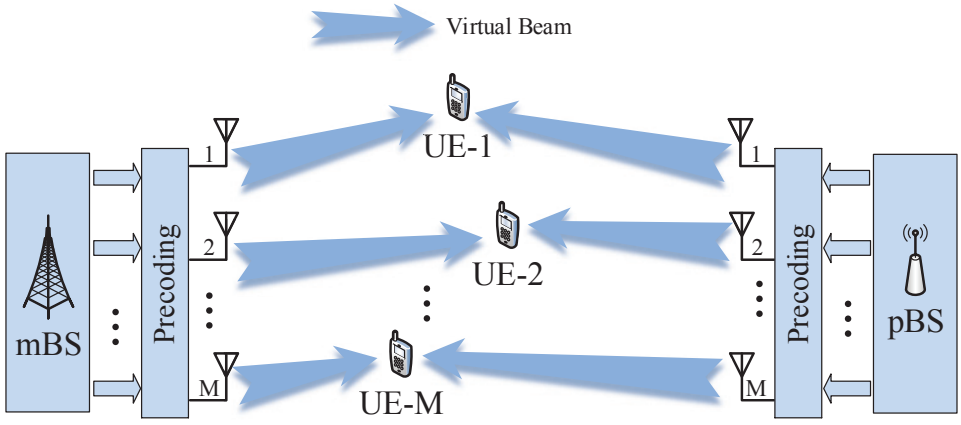

Fig. 5.2: Transmission channel model with DPC

5.3.1 Dirty Paper Coding

Dirty paper coding is a precoding technique to cancel out inter-user interference on the transmitting side [33]. In particular, we assume DPC precoding matrix is designed by given perfect channel state information. Without loss of generality, we can formulate the channel

matrices $\mathbf{H}_i$ and $\mathbf{H}_j$ as:

$$\mathbf{H}_i = \begin{bmatrix} h_{i,1,1} & \cdots & h_{i,1,M} \\ \vdots & \ddots & \vdots \\ h_{i,M,1} & \cdots & h_{i,M,M} \end{bmatrix}_{M \times M}, \tag{5.2}$$

and

$$\mathbf{H}_j = \begin{bmatrix} h_{j,1,1} & \cdots & h_{j,1,M} \\ \vdots & \ddots & \vdots \\ h_{j,M,1} & \cdots & h_{j,M,M} \end{bmatrix}_{M \times M}, \tag{5.3}$$

where $h_{i(j),\tilde{m},m_k}$ denotes the channel gain between UE k with index m_k in the cooperative set and the $\tilde{m}$th transmitting antenna of mBS i (pBS (i,j)). This channel gain considers both large scale and small scale fading. For the sake of simplicity, we omit the subscripts and assume $\mathbf{H} \in \{\mathbf{H}_i, \mathbf{H}_j\}$. By applying LQ decomposition to $\mathbf{H}$, we can obtain:

$$\mathbf{H} = \begin{bmatrix} l_{1,1} & 0 & \cdots & 0 \\ l_{2,1} & l_{2,2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ l_{M,1} & l_{M,2} & \cdots & l_{M,M} \end{bmatrix}_{M \times M} \begin{bmatrix} \mathbf{q_1} \\ \mathbf{q_2} \\ \vdots \\ \mathbf{q_M} \end{bmatrix}_{M \times M} = \mathbf{LQ}. \tag{5.4}$$

Here, $\mathbf{L}$ is an $M \times M$ lower triangle matrix and $\mathbf{Q}$ is an $M \times M$ semi-unitary matrix. As a result, the precoding matrix $\mathbf{W}$ can be expressed as

$$\mathbf{W} = \mathbf{Q}^H \mathbf{G}, \tag{5.5}$$

in which $\mathbf{Q}^H$ is the Hermitian matrix of $\mathbf{Q}$ and $\mathbf{G}$ should satisfy the following criteria so that we can obtain an ideally interference-free transmission channel:

$$\mathbf{G} = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ -\frac{l_{1,2}}{l_{2,2}} & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ -\sum_{i=1}^{M-1} \frac{l_{M,i}}{l_{M,M}} \mathbf{g}_{i,1} & -\sum_{i=1}^{M-1} \frac{l_{M,i}}{l_{M,M}} \mathbf{g}_{i,2} & \cdots & 1 \end{bmatrix}_{M \times M}. \tag{5.6}$$

$\mathbf{G}$ is also an $M \times M$ lower triangle matrix whose entry $\mathbf{g}_{i,j}$ at ith row and jth column has the following expression:

$$\mathbf{g}_{i,j} = -\sum_{i'=1}^{i-1} \frac{l_{i,i'}}{l_{i,i}} \mathbf{g}_{i',j}. \tag{5.7}$$

By transmitting the precoded signal $\tilde{\mathbf{x}} = \mathbf{W}\mathbf{x}$ for $\mathbf{x} \in \{\mathbf{x}_i, \mathbf{x}_j\}$, the received signal is given by

$$\begin{aligned} \mathbf{y}_{i,j,\mathcal{K}_{i,j}^f(t)} &= \mathbf{H}_i \tilde{\mathbf{x}}_i + \mathbf{H}_j \tilde{\mathbf{x}}_j + \mathbf{z} = \mathbf{H}_i \mathbf{W}_i \mathbf{x}_i + \mathbf{H}_j \mathbf{W}_j \mathbf{x}_j + \mathbf{z} \\ &= \mathbf{L}_i \mathbf{Q}_i \mathbf{Q}_i^H \mathbf{G}_i \mathbf{x}_i + \mathbf{L}_j \mathbf{Q}_j \mathbf{Q}_j^H \mathbf{G}_j \mathbf{x}_j + \mathbf{z} \\ &= \begin{bmatrix} l_{i,1,1} & 0 & \cdots & 0 \\ 0 & l_{i,2,2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & l_{i,M,M} \end{bmatrix}_{M \times M} \begin{bmatrix} x_{i,1} \\ x_{i,2} \\ \vdots \\ x_{i,M} \end{bmatrix} \\ &\quad + \begin{bmatrix} l_{j,1,1} & 0 & \cdots & 0 \\ 0 & l_{j,2,2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & l_{j,M,M} \end{bmatrix}_{M \times M} \begin{bmatrix} x_{j,1} \\ x_{j,2} \\ \vdots \\ x_{j,M} \end{bmatrix} + \begin{bmatrix} z_1 \\ z_2 \\ \vdots \\ z_M \end{bmatrix} \\ &= \mathbf{D}_i \mathbf{x}_i + \mathbf{D}_j \mathbf{x}_j + \mathbf{z}. \end{aligned} \tag{5.8}$$

It is observed that the inter-user interference is eliminated ideally assuming we have perfect channel knowledge. As shown in Fig. 5.2, after precoding, mBS/pBS transmitter generates

multiple beams like an MU-MIMO. Two beams, one from an mBS and one from a pBS, are superimposed at a UE. In the DPC+NOMA scheme, each multi-antenna BS can cancel out inter-user interference by using DPC while each UE receives an aggregated signal from multiple BSs if NOMA is applied. Next section will give more details on how SIC scheme decodes the composite signals at UEs.

5.3.2 Non-orthogonal Multiple Access with Successive Interference Cancellation

NOMA enables different signals to be transmitted over the same radio resource frequency to the same receiver at the same time. In our scheme, an mBS and a pBS can transmit different signals to the same UE simultaneously. A UE will be able to extract the received signal from the respective BS by using SIC algorithm, which can decode the signal from each BS sequentially, in the descending order of the received signal strength. Given the effective channel matrices $\mathbf{D}_i$ and $\mathbf{D}_j$, a UE will first decode the stronger signal received from either mBS ($\mathbf{x}_i$) or pBS ($\mathbf{x}_j$) by treating other weaker signals as interference, and then subtracts the decoded signal from $\mathbf{y}_{i,j,\mathcal{K}_{i,j}^f(t)}$. UE repeats the decoding process until all the signals are successfully decoded. To realize the most efficient NOMA gain, the signals received from different BSs should have evident power difference so that at each SIC iteration, a high effective SINR is achieved. To that end, we exploit the power disparity among mBSs and pBSs and use different mobile association schemes to maximize the number of candidate UEs for NOMA. The throughput for a UE receiving from mBS i and pBS(i,j) is given by

$$\mathbf{R}_{i,j} = W \log_2 \left(\mathbf{I} + \frac{\hat{\mathbf{D}} \hat{P} \hat{\mathbf{D}}^H}{\tilde{\mathbf{D}} \tilde{P} \tilde{\mathbf{D}}^H + \mathbf{z}} \right) + W \log_2 \left(\mathbf{I} + \frac{\tilde{\mathbf{D}} \tilde{P} \tilde{\mathbf{D}}^H}{\mathbf{z}} \right). \tag{5.9}$$

Here, $\mathbf{R}_{i,j}$ is an $M \times M$ diagonal matrix in which the mth diagonal entry represents the throughout at UE k with a local index m, $1 \leq m \leq M$. W is the bandwidth per RB, and

$$\hat{\mathbf{D}} \hat{P} \hat{\mathbf{D}}^H = \max \left(\mathbf{D}_i P_m \mathbf{D}_i^H, \mathbf{D}_j P_p \mathbf{D}_j^H \right), \tag{5.10}$$

$$\tilde{\mathbf{D}} \tilde{P} \tilde{\mathbf{D}}^H = \min \left(\mathbf{D}_i P_m \mathbf{D}_i^H, \mathbf{D}_j P_p \mathbf{D}_j^H \right). \tag{5.11}$$

5.4 Problem Formulation

We aim to implement the DPC + NOMA scheme so that the overall system throughput can be maximized. To that end, at each scheduling cycle and for each RB, we should:

(1) form UE cluster for cooperation for all BSs;

(2) for each UE within an identified cluster, use NOMA to let mBS i and pBS j transmit different data to this UE.

The algorithm will determine UE cluster and serving mBS and pBS altogether for each RB. Once the pair (mBS i, pBS (i,j)) is determined, the rest pBSs (i,j'), $\forall j' \neq j$, will switch to muting mode so that there is no intra-cell interference. Due to cooperation, multiple UEs within the same cooperation cluster will be served by a multi-antenna BS. Thus the following scheduling variables are defined.

$$x_{i,0,m_1(k)}(f,t) = \begin{cases} 1; & \text{UE } k \text{ is served by mBS } i \text{ on } f \text{ as the } m_1\text{th UE in a cluster at } t, \\ 0; & \text{otherwise.} \end{cases}$$

$$x_{i,j,m_2(k)}(f,t) = \begin{cases} 1; & \text{UE } k \text{ is served by pBS } (i,j) \text{ on } f \text{ as the } m_2\text{th UE in a cluster at } t, \\ 0; & \text{otherwise.} \end{cases}$$

Furthermore, we define $\mathbf{R}_k(t)$ as the window-based moving average throughput for UE k at time t as:

$$\mathbf{R}_k(t) = \frac{1}{T_c} \sum_{\tau=t-T_c+1}^{t} S_k(\tau), \tag{5.12}$$

where T_c is the moving average window size. $S_k(\tau)$ is the instantaneous throughput for UE k at τ, which is expressed as

$$S_k(\tau) = \sum_{i=1}^{N_m} \sum_{j=1}^{N_p} \sum_{f=1}^{F} x_{i,0,m_1(k)}(f,\tau) \times x_{i,j,m_2(k)}(f,\tau) \times \mathbf{R}_{i,j,m_3(k)}(f,\tau), \tag{5.13}$$

for $m_i = 1, \cdots, M$. Here, $\mathbf{R}_{i,j,m_3(k)}(f,t)$ is the m_3th entry in $\mathbf{R}_{i,j}$ obtained from in (5.9), which means at each scheduling circle, data rate of UE k corresponds to the m_3th entry in $\mathbf{R}_{i,j}$. The objective function of the scheduling problem is thus formulated as

$$\mathbf{P}_1 : \max_{\mathbf{x}(t)} \sum_k \mathbf{R}_k(t) \tag{5.14}$$

subject to

$$\sum_{m_1=1}^{M} \sum_{k=1}^{N_u} x_{i,0,m_1(k)}(f,t) \leq M, \ \forall i, f, m, \tag{5.15}$$

$$\sum_{m_2=1}^{M} \sum_{k=1}^{N_u} x_{i,j,m_2(k)}(f,t) \leq M, \ \forall i, j, f, m, \tag{5.16}$$

where (5.15) and (5.16) indicate that at most M UEs can share the same RB f from mBS i and pBS (i,j) at time t.

In order to solve the problem, a cooperative cluster $\mathcal{K}_{i,j}^f(t)$ needs to be decided at each scheduling circle, where $\mathcal{K}_{i,j}^f(t)$ has a cardinality M. RB f should be assigned to UEs in this set ($k \in \mathcal{K}_{i,j}^f(t)$) so that the instantaneous throughput achieves maximum. Therefore, the objective function is equivalent to:

$$\mathbf{P}_2 : \max_{\mathbf{x}(t)} \sum_k \mathbf{S}_k(t), \forall t \tag{5.17}$$

In the following section, we introduce a genetic algorithm to achieve trade-off between optimality and computational efficiency.

5.5 Genetic Algorithm

In this section, we apply a genetic algorithm (GA) [35] to search the best UE clusters at time t. Compared with the traditional brute-force search, genetic algorithm can greatly reduce computational workload but still retain good performance.

In order to narrow down the search space and decrease computational complexity, we only consider UE cluster in the same sector. In the ith sector, we can form a clustering

candidacy set $\mathcal{K}_i$ and choose clustering UEs from this set. To start genetic algorithm, we first need to encode a possible solution as a string of bits:

$$\mathcal{C}_i : \quad [\text{mBS}]\ [\text{pBS}]\ [\text{UE } 1]\ \cdots\ [\text{UE } M], \tag{5.18}$$

where [mBS] is the binary sequence for mBS's index, [pBS] is the binary sequence for pBS's index, [UE k_1] is the binary sequence for 1st clustering UE's index, and [UE k_M] is the binary sequence for Mth clustering UE's index. These sequences can be regarded as genes, which are all strung together to form a chromosome. Once the chromosome is determined, we can decode the chromosome and obtain the clustering UEs as well as their association status.

After encoding, GA will proceed to initialize a population of possible solutions. Through repetitive application of genetic operators and being filtered by fitness function, the possible solutions are getting improved. The process is repeated until a proposed termination condition has been reached. Overall, GA is proceeded in four steps.

1. **Generate population of solutions:**

   Based on the defined genetic representation in (5.18), we randomly generate N possible solutions.

2. **Selection:**

   We evaluate the possible solution through a fitness-based process, where the fitter solutions are typically more likely to be selected. In this paper, fitter solution is measured by the fitness function which is defined in (5.9). That is, we keep the possible user cluster that achieves highest aggregated throughput.

3. **Genetic operation:**

   After selection, the next step is to generate a second generation population of solutions, with population size N, from the fitter solutions. Mutation is one of genetic operations to generate a new possible solution. In each generation, mutation can occur at pBS's index sequence, and the clustering UE's index sequences. Usually, each bit in the

sequence has a mutation probability. In our case, bit with the highest mutation probability will be changed from its original value, and the resultant chromosome is a newly generated possible solution.

4. **Termination:**

   We evaluate the new generation of possible solutions through fitness function, and find the best UE cluster that contributes the highest aggregated throughput. Then we check if the GA process is terminated by the following rules:

   (a) drift value between the newly generated solution and parent solution is less than convergence threshold ϵ;

   (b) fixed number of generations reached.

   If one of the above two conditions is satisfied, the GA process is terminated. Otherwise, the generational process repeats from step (b).

For additional clarity, we summarize the genetic algorithm in Algorithm 2.

5.6 Performance Evaluation

The simulation was set up based on 3GPP case 1 configurations specified in [22]. In a 19-cell 3-sector three-ring hexagonal network structure, one mBS is located in the center of a macro-cell and 4 pBSs are equally-distanced deployed in the overlaid pico-cells within each macro-cell, forming a two-tier heterogeneous network. UEs are uniformly distributed in the network. Fast fading is generated based on the Rayleigh fading channel model in [34]. In this paper, we assume the cluster size $M = 2$. Other parameter settings are shown in Table 5.1.

For optimality comparison, we introduce a greedy user pairing algorithm ("Greedy" for short), which is very similar to brute-force search. The algorithm is inspired by [36] and is divided into two phases: 1) In the first phase, the system finds out the best UE k_1 which can achieve the highest value of objective function; 2) In the second phase, search the best pairing UE k_2 so that the aggregated data rate of k_1 and k_2 (objective function value) is

Algorithm 2 Genetic Algorithm

1: **Initialization**: chromosome $\mathcal{C}_i$ for sector i ($1 \leq i \leq N_m$), $\epsilon > 0$ as the convergence threshold, g is the generation index and G as the maximum number of generations. Fitness function is based on (5.9) and denoted as $F(\mathcal{C}_i)$.
2: **for** $f = 1$ to F **do**
3: **for** $i = 1$ to N_m **do**
4: Randomly choose one pBS in sector i and two UEs from set $\mathcal{K}_i$. Generate a solution set $\mathcal{S}_i(g)$ which is consisted of N unique chromosomes ($g = 0$).
5: **repeat**
6: **Select** the fitter solution from set $\mathcal{S}_i(g)$ to maximize the fitness function:

$$\mathcal{C}_i^*(g) = \arg\max_{\mathcal{S}_i(g)} F(\mathcal{C}_i(g)) \tag{5.19}$$

7: **Mutate** $\mathcal{C}_i^*(g)$ to produce next generation solution set $\mathcal{S}_i(g+1)$ which is consisted of N new chromosomes:

$$\mathcal{C}_i^*(g) \rightarrow \mathcal{C}_i^1(g+1) \cdots \mathcal{C}_i^N(g+1) \tag{5.20}$$

8: $g = g + 1$
9: **until**
10: $F(\mathcal{C}_i^*(g)) - F(\mathcal{C}_i^*(g-1)) \leq \epsilon$, or
11: $g \geq G$
12: **end for**
13: **Decode** chromosome $\mathcal{C}_i^*$, find out the best UE cluster and serving pBS

$$\mathcal{K}_{i,j^*}^f = \{k_1^*, \cdots k_M^*\} \in \mathcal{K}_{i^*}. \tag{5.21}$$

14: Allocate RB f to UE cluster $\{k_1^*, \cdots k_M^*\}$
15: **end for**
16: Output F optimal user clusters $\{k_1^*, \cdots k_M^*\}$, indexes of corresponding serving mBS and pBS (i, j^*) and resource scheduling decisions at each time slot.

maximum. Besides, we also compare our proposed cooperative NOMA scheme with the other two schemes:

(1) DPC cooperation only scheme ("DPC" for short): A single mBS or pBS with 2 transmitting antennas communicates with two UEs;

(2) NOMA-SIC only scheme ("NOMA-SIC" for short): An mBS with a single antenna and a pBS with a single antenna jointly communicate with a UE by using NOMA.

It is noted that with DPC scheme, each mBS or pBS is equipped with two antenna and it can serve two UEs simultaneously. Whereas with NOMA-SIC scheme, mBS and pBS

Table 5.1: Parameter settings

Parameter	**Settings**
mBS	57
pBS	4 per macro-cell
UE	50 per cell
Number of Transmitting Antenna	2 per BS
Number of Receiving Antenna	1 per UE
Size of Cluster	$M = 2$
Transmit Power	$P_m = 30$Watt, $P_p = 1$Watt
System Bandwidth	10 MHz
Number of RBs	$F = 50$
Bandwidth of RB	$W = 180$kHz
Size of Time Window	$T_c = 1$ second
Fast Fading Model	Rayleigh Fading Channel [34]
Path loss from mBS to UE	$LOS(d) = 103.4 + 24.2\log_{10}(d)$ $NLOS(d) = 131.1 + 42.8\log_{10}(d)$
Path loss from pBS to UE	$LOS(d) = 103.8 + 20.9\log_{10}(d)$ $NLOS(d) = 145.4 + 37.5\log_{10}(d)$
Shadowing	8dB, log-normal std. deviation
Noise Model and density	AWGN, -174dBm/Hz
Chromosome Format	aaaaaa bb cccccc dddddd
Population Size	$N = 200$
Maximum Number of Generations	$G = 24$
Convergence Threshold	$\epsilon = 10^{-2}$

are equipped with single transmitting antenna, respectively. They transmit two different signals to one UE on the same frequency band. Without loss of generality, both of above two schemes are also implemented by using GA.

In each generation, we evaluate the network throughput as the relative percentage of the throughput of greedy algorithm, which can be regarded as the ideal optimal solution. In

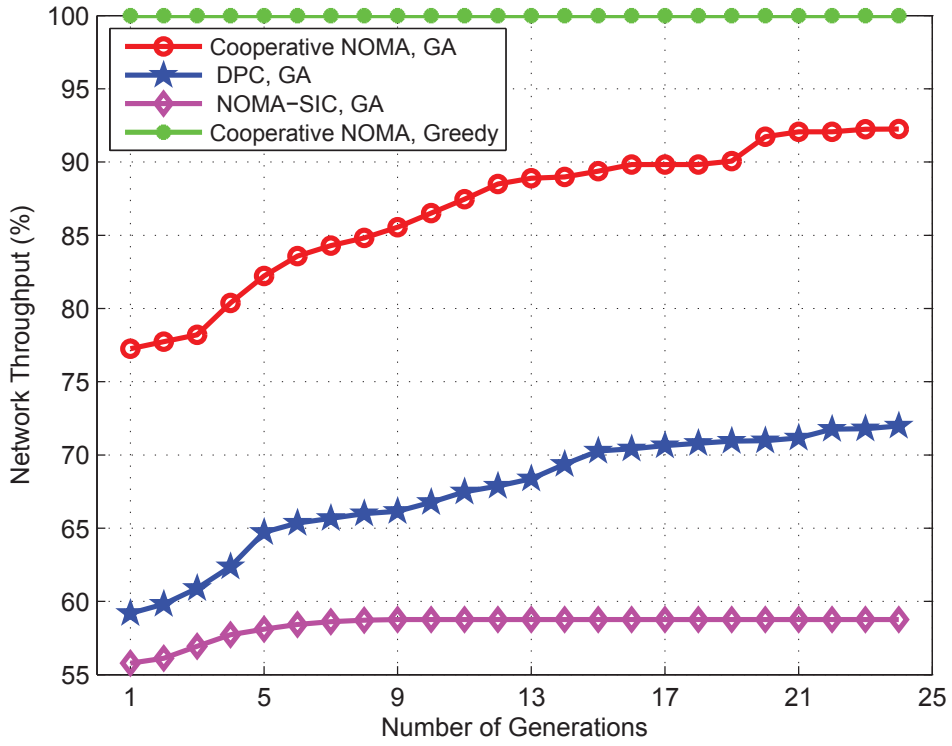

Fig. 5.3: Network throughput under different schemes, $N = 200$

Fig. 5.3, it is observed that the system instantaneous throughput goes up with the increment of generations. Specifically, the proposed cooperative NOMA scheme always has a higher throughput than that from the other two schemes. It can reach to over 92% of the optimal value. In contrast, DPC can achieve at most 72% and NOMA-SIC can only achieve 58%. This is because by applying cooperative NOMA scheme, one mBS and one pBS can serve two UEs simultaneously, and each UE can receive two different signals at the same time. In other word, cooperative NOMA scheme enables two BSs to serve more UEs than NOMA-SIC scheme, and enables UEs to receive more signals than DPC scheme. Therefore, the proposed scheme can achieve the best performance gain.

In Fig. 5.4, we study GA performance trade-off between computational complexity and optimality under cooperative NOMA scheme. We evaluate the performance gain with respect to different population sizes $N = \{100, 200, 400\}$. It is observed that with the increment of the population size, the achievable performance gets closer to the ideal optimal value, which means the gap between GA and greedy algorithm is reduced. The reason is

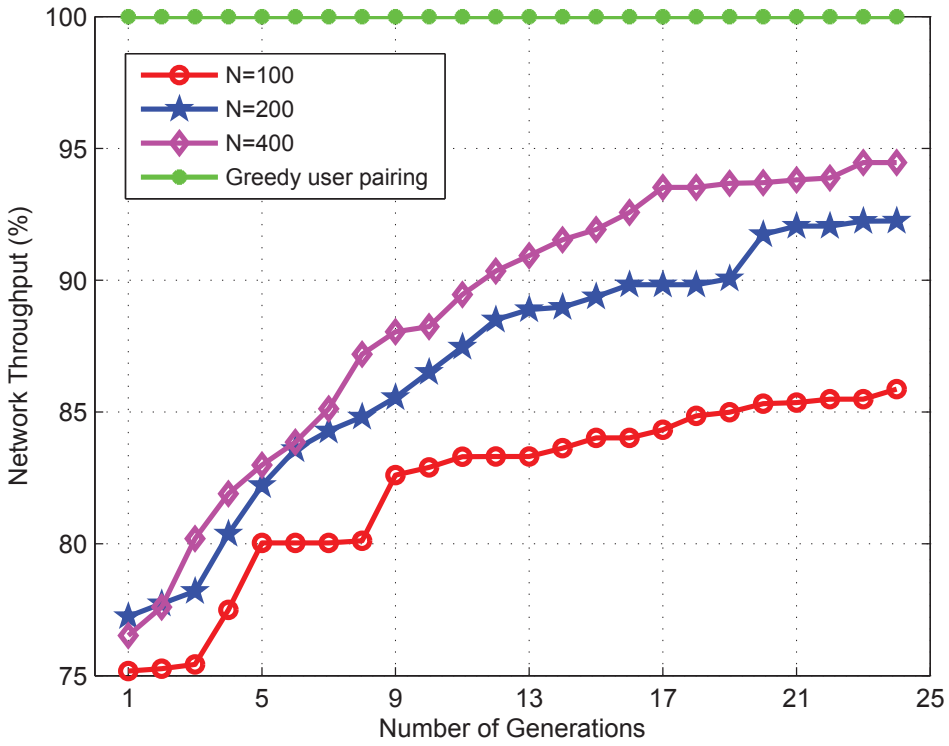

Fig. 5.4: Network throughput under different population size N

that as a heuristic algorithm, GA contains certain randomness. With a larger population size, it has a higher chance to find out the best UE cluster in each generation. Thus, the result is much closer to the ideal optimal value. Nevertheless, a larger population size N leads to a higher computational complexity. For example, the result for $N = 400$ is about 94%, and the result for $N = 200$ is about 92%. Although we can obtain a 2% optimality gain by choosing $N = 400$, we almost double the computational workload in each generation. Consequently, $N = 200$ is a reasonable trade-off between computation complexity and optimality gain.

5.7 Chapter Summary

In this chapter, we investigate a cooperative NOMA scheme in a downlink heterogeneous network, by integrating DPC and NOMA together to increase the transmission rates and achieve cooperation gain. In order to allocate resources to the best cooperative users, we introduce a genetic algorithm to balance the trade-off between computational complex-

ity and optimality. Extensive simulations are conducted to compare the proposed scheme with cooperation scheme and conventional NOMA-SIC scheme. Besides, we also study the performance trade-off of GA between computational complexity and optimality. In the future, we would like to extend our work to a uplink transmission scenario by considering imperfect SIC process with error propagation in information recovery.

Chapter 6

Video Quality-based Spectrum and Energy Efficient Mobile Association in Wireless Heterogeneous Networks

6.1 Introduction

The explosive growth in mobile video applications, enabled by a plethora of powerful handset devices, brought to the forefront the dramatic increase in spectrum demand and energy consumption in wireless networks. Mobile devices, such as smart-phones and tablets, are increasingly being used in social environments for video chatting, video streaming, and movie downloading. According to [37], in year 2012, mobile video traffic exceeded, for the first time, 50% of the total wireless traffic. Furthermore, it is expected that mobile video will increase 16-fold between 2012 and 2017, to account for over 66% of the total mobile data traffic, by the end of 2017.

The expected increase of data generated by video applications comes at the high price of exponential increase in energy and spectrum consumptions. In traditional cellular networks, a BS consumes a significant amount of energy to support the activities of UEs, especially the cell edge users. Emerging highly-dense, wireless heterogeneous networks introduce a hierarchical infrastructure, where high power BSs provide blanket coverage and seamless mobility, while low power nodes, such as femto- and pico-BS, help support cell edge users and boost cell capacity [1–3]. Usually deployed at coverage holes or capacity-demanding hotspots, these low power nodes can extend the wireless service coverage range and expand the cell capacity.

In previous chapters, we discussed the optimal mobile association and resource allocation schemes in wireless heterogeneous networks, with the application of promising technologies such as CoMP, DPC and NOMA. The aforementioned work, however, mainly focuses on

traditional data transmission. Moreover, existing heterogeneous network mobile association schemes are mainly based on system spectrum efficiency (SE) or energy efficiency (EE), in which every bit transmission contributes equally to the system SE and EE. As such, it does not consider UEs' video quality requirements and their impact on SE and EE, in mobile association. In [38–40], the authors investigate energy management for multi-homing video transmission in heterogeneous networks. In these works, the network heterogeneity is not intrinsic to the BSs, but rather limited to differences in services interfaces. These limitations underscore the need to explore the relationship between the system's SE and EE and their impact on video applications' quality requirements, in emerging wireless networks.

The main objective of this chapter is to address the shortcomings of current paradigms and explore mobile association and resource allocation in wireless heterogeneous networks, emphasizing the interplay between video quality and resource consumption. To this end, we propose two new system performance metrics, QSE and QEE, which measure the video quality per unit radio resource consumption and per unit power consumption, respectively. The two performance metrics can be viewed as an extension of the traditional SE and EE metrics. Given that not all frames of a video content are of equal importance and delivery priority, the video peak-signal-noise-ratio (PSNR) is used to characterize video quality. Based on this characterization, QSE and QEE are defined as PSNR/Hz and PSNR/Watt, respectively. The fundamental trade-off between QSE and QEE is studied in a point-to-point (PtP) additive white Gaussian noise (AWGN) wireless channel and Rayleigh fading channel, respectively.

Based on the QSE and QEE metrics, we formulate a joint mobile association and resource allocation optimization problem in a heterogeneous network. The problem is initially a mixed integer nonlinear programming problem (MINLP) with fractional form that cannot be solved in polynomial time. Hence, we apply Dinkelbach's method in nonlinear fractional programming [41]. We then use both linear relaxation and variable transformation to reduce the original optimization problem to a computationally tractable problem. Furthermore, in order to achieve a good balance between the computational complexity and optimality, we

introduce Lagrange dual decomposition to decompose the problem into a sequence of similar sub-problems, which can be solved parallelly. At the core of the QSE and QEE based mobile association problem is a decision to accept or reject a mobile user's request to establish a video connection. If the request is accepted, the system must also determine the amount of radio resources to be reserved for new connection, in order to maximize overall system QSE and QEE. Note that these decisions must be made before the connection is established. Furthermore, we undertake an advanced system-level study to explore the trade-off study between QSE and QEE. To this end, we formulate a multi-objective optimization problem (MOOP) to investigate the trade-off and interplay between QSE and QEE, in a wireless heterogeneous network. It is worth noting that the two cases investigated in [42], represent special instances of the MOOP problem, which can be derived by setting one of the weights of the objective functions, ω_1 or ω_2 to zero.

In the following sections, we use boldface capital and boldface lower case letters to represent matrices and vectors, respectively. $\mathbf{I}$ is denoted as an all-one matrix. $\mathbf{1}$ is denoted as an all-one column vector. $\mathbf{A} \circ \mathbf{B}$ represents the Hadamard product of matrix $\mathbf{A}$ and $\mathbf{B}$. $Diag\{\mathbf{A}\}$ is denoted as the diagonal column vector of matrix $\mathbf{A}$. $Tr\{\mathbf{A}\}$ is denoted as the trace of matrix $\mathbf{A}$. $\mathbb{E}\{f(x)\}$ represents the ensemble average of the function $f(x)$ over the probability density function of the random variable x.

6.2 Video Content Delivery over Heterogeneous Wireless Networks

We consider a two-tier heterogeneous wireless network shown in Fig. 5.1, where each macro-cell hosts one mBS and several overlaid pBSs. Both mBS and pBS are connected with the video server via wired links. In particular, we denote the number of mBS as N_c, the number of pBS per macro-cell as N_r, and the number of UEs as N_u. Thus, the total number of pBSs is denoted as $N_p = N_c \times N_r$.

6.2.1 Video Quality Measurement

Our study will focus on the mobile association for video applications in heterogeneous networks. In order to most effectively attach video mobiles to the right serving BSs that

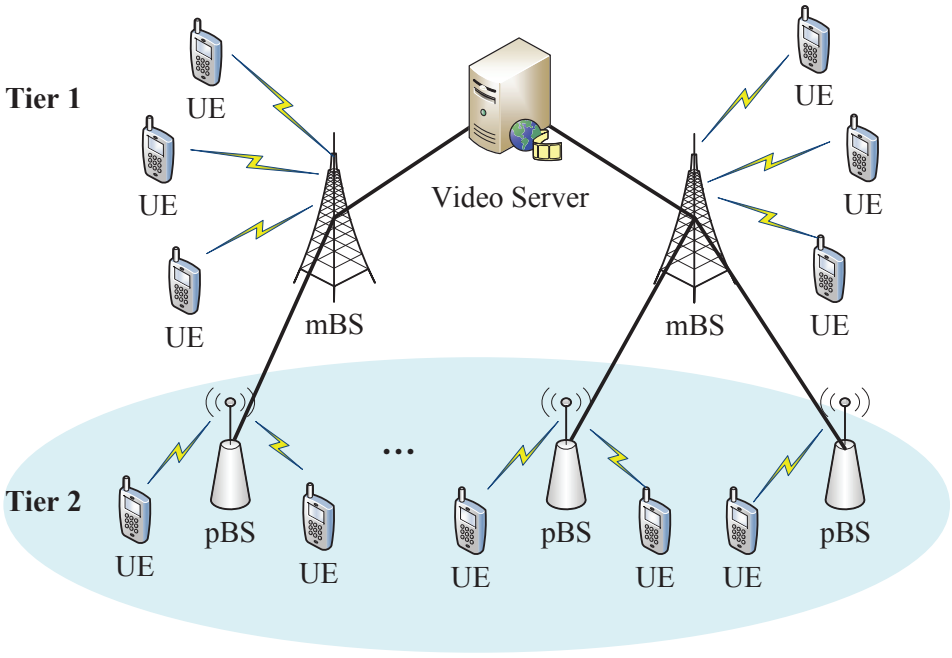

Fig. 6.1: Two-tier wireless heterogeneous network model

can spectrum/energy efficiently deliver the required video quality, we need to consider SE, EE and video quality altogether during mobile association. PSNR is commonly used as a metric to measure objective video quality. Objective video models are mathematical models that approximate results from subjective quality assessment, in which human observers are asked to rate the quality of a video. In this paper, we use the definition in [43], where PSNR is approximated as a logarithmic function of data rate:

$$\text{PSNR} = \alpha \log_{10}(R), \tag{6.1}$$

where α is a predefined parameter that is related to video contents and R is the achievable data rate over wireless channels. We capture video quality requirements in the mobile association process for the decision making. Mobile association is similar to the traditional call admission control process, where there is no video traffic flowing in the network yet. Different users may request different video contents, which may have different α values and thus different quality requirements.

6.2.2 Video Quality-aware Spectrum Efficiency and Energy Efficiency

In modern wireless communications, system design mainly focuses on achieving desirable SE and EE. A large body research work has focused on achieving this goal. Recent advances in video communications ushered in new opportunities, but also brought about new challenges. In the traditional wireless network design, each bit transmission contributes equally to the network throughput, which usually scales linearly with SE and EE. Video quality differentiation exposes the inadequacy of the conventional SE and EE metrics to reflect the design requirements of video applications, in which video quality does not linearly scale with its data rate or throughput. The increase of video quality tends to saturate when the data rate exceeds certain level. This, in turn, underscores the necessity to analyze SE and EE from a video quality's perspective. In this paper, we would like to design a QSE and QEE based mobile association that aims to attain the best trade-off among SE, EE and video quality.

Without loss of generality, we denote the total bandwidth consumption as W and the total power consumption as P. Thus QSE and QEE are defined as

$$\text{QSE} \triangleq \frac{\text{PSNR}}{W^{\theta}} \quad \text{and} \quad \text{QEE} \triangleq \frac{\text{PSNR}}{P^{\beta}}, \tag{6.2}$$

where θ and β are the decaying factors that respectively indicate the relative costs of the bandwidth consumption and power consumption when delivering PSNR. Traditional models usually use flat rates in resource pricing and formulate the cost function as a linear relationship of the consumed resources [44]. In this work, we consider a more realistic resource pricing model, which characterizes both the consumer's behavior and supplier's business strategies. In real environments, suppliers (e.g., National Grid, Comcast, and Verizon) usually provide considerable discounts to their clients to encourage resource consumption and consequently increase their revenue. On the other hand, clients, such as technology companies and financial corporations, need to consume a large amount of resources to keep their business running. They seek significant discounts from their suppliers so they can control their operation costs. To address clients' requirements, while remaining profitable, suppliers

apply a usage-based linear pricing model if the resource consumption is relatively small. For large resource consumption, suppliers apply a relatively flat and slowly increasing pricing model, to encourage clients to consume more resources. This pricing model is depicted in Fig. 6.2. It can be observed that in a large resource consumption region, the price increases relatively smoothly. The blue curve shows the linear relationship between price and resource consumption, which is widely used in most of the existing research works. Although simple and straightforward, the model fails to consider consumer' behaviors and realistic company business strategies. Furthermore, a higher θ or β value means that wireless network is less concerned about video quality PSNR, but more concerned about bandwidth or power consumption. Thus the decaying factors in QSE and QEE aim to strike a balance between the gain in PSNR value and the cost of bandwidth or power consumption.

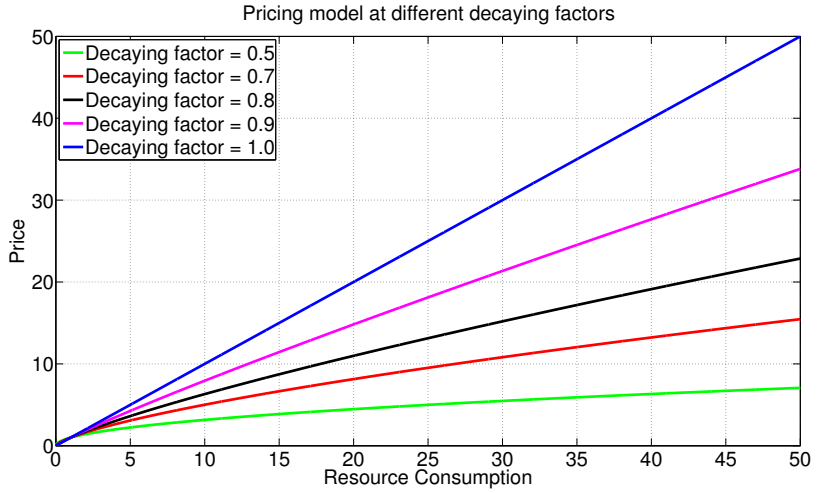

Fig. 6.2: Network resource pricing model

6.3 QSE and QEE in PtP AWGN Channel

Motivated by [45], we aim to establish the fundamental relationship between QSE and QEE in a PtP AWGN channel. Furthermore, we would like to use the same power model that has been used in the earlier fundamental study on SE and EE trade-off in [45] for comparison purpose. Thereby, we assume the total power consumption P consists of two parts: dynamic part which is mainly transmit power P_t and static part which is mainly from

circuit power P_c. According to Shannon formula [46], we have $R = W \log_2\left(1 + \frac{P_t h}{W N_0}\right)$, where W is the allocated bandwidth, h is the channel gain and N_0 is the AWGN noise density. Then we have:

$$\text{QSE} = \frac{\alpha \log_{10}\left[W \log_2\left(1 + \frac{P_t h}{W N_0}\right)\right]}{W^{\theta}}, \tag{6.3}$$

$$\text{QEE} = \frac{\alpha \log_{10}\left[W \log_2\left(1 + \frac{P_t h}{W N_0}\right)\right]}{(P_t + P_c)^{\beta}}. \tag{6.4}$$

As PSNR is a logarithm function of the video throughput, any further increase for an already high video throughput will only lead to a marginal increase on the PSNR value. This is consistent with the understanding that base layer with the lowest data rate contributes the most to PSNR, and provides the fundamental information of video contents. The higher the video frame layer, the smaller PSNR the frame contributes. This is because the reception of enhancement video layers usually refines the details of video contents. Thus, it makes the bandwidth and power consumptions video quality-aware possible.

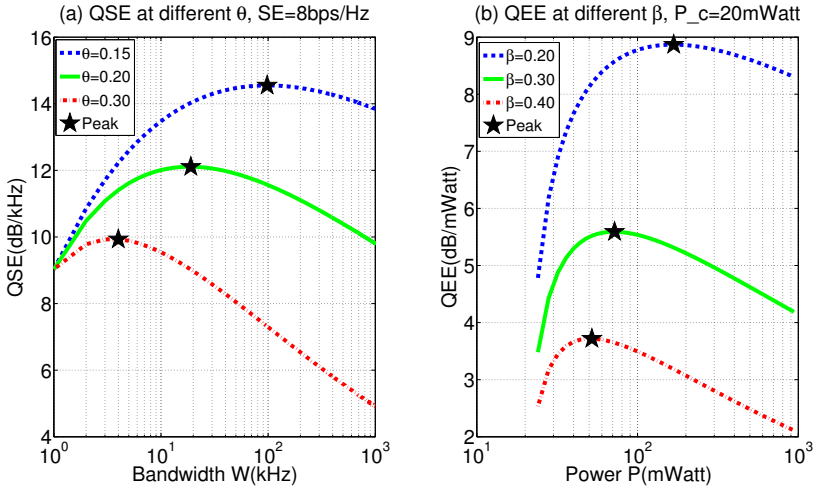

Fig. 6.3: QSE/QEE performance at different decaying factors

Fig. 6.3 illustrates the impact of the decaying factor on the QSE and QEE performance, respectively. Specifically, Fig. 6.3(a) shows QSE-W curves for different values of θ and Fig. 6.3(b) shows QEE-P curves for different values of β. When the value of θ (β) is larger,

QSE (QEE) achieves its peak value at a smaller W (P) and at a lower PSNR, and thus at a lower peak QSE (QEE). Since a larger decaying factor tends to be more stingy on bandwidth or power consumption, then the PSNR is low. When θ or β decreases, the cost of bandwidth or power consumption also decreases. Correspondingly, the system is more willing to achieve a relatively higher PSNR with a higher bandwidth or power consumption. In the following, we explain why QSE/QEE is a bell-shaped curve with respect to W/P.

In the low bandwidth/power region, the increment of PSNR is faster than the increment of bandwidth or power consumption, so that QSE-W or QEE-P curve goes up. After QSE-W or QEE-P each reaches respective peak value, the increment of bandwidth or power consumption surpasses the increment of PSNR so that the QSE-W or QEE-P curve goes down. In the high bandwidth/power region, we notice that both QSE-W and QEE-P curves become flat, when the decaying factors are small. Recall that a smaller decaying factor indicates a lower cost on W or P. Thus, with a small decaying factor, both PSNR and W^θ (P^β) increase very slowly in the high W^θ (P^β) region, making the QSE-W and QEE-P curves flat.

Furthermore, from the expression of QSE in (6.3), P_t can be expressed as a function of W and QSE:

$$P_t = h^{-1} W N_0 \left[2^{W^{-1} 10^{(\alpha^{-1} \mathrm{QSE} * W^\theta)}} - 1 \right]. \tag{6.5}$$

By inserting P_t into the formula in QEE, the trade-off between QSE and QEE can be expressed as:

$$\mathrm{QEE} = \frac{\mathrm{QSE} * W^\theta}{\left(h^{-1} W N_0 \left[2^{W^{-1} 10^{(\alpha^{-1} \mathrm{QSE} * W^\theta)}} - 1 \right] P_c \right)^\beta}. \tag{6.6}$$

Assuming $\theta = \beta = 1$ and $P_c = 0$, and replacing PSNR by R, QSE and QEE reduce to the traditional SE and EE expressions [45]:

$$SE = \frac{R}{W} = \log_2 \left(1 + \frac{P_t h}{W N_0} \right), \tag{6.7}$$

$$EE = \frac{R}{P} = \frac{SE}{h^{-1} N_0 \left(2^{SE} - 1 \right)}. \tag{6.8}$$

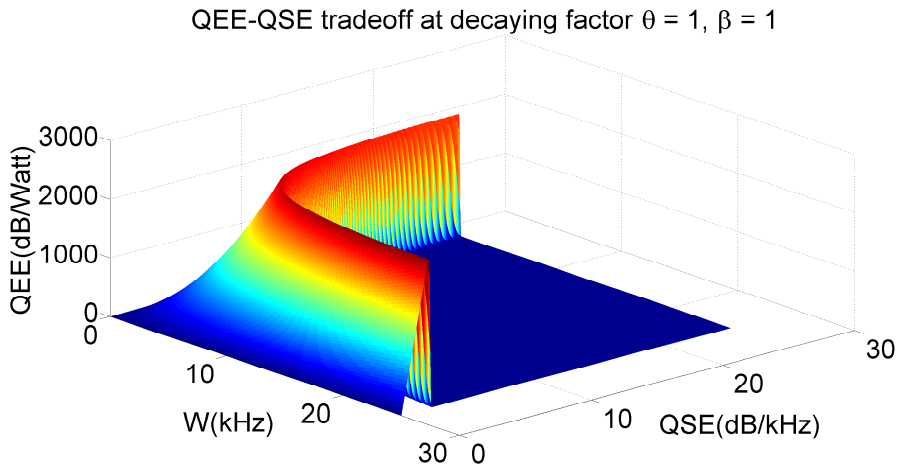

Fig. 6.4: QSE-QEE trade-off at decaying factors $\theta = 1$, $\beta = 1$

From (6.6), we can observe that QEE is not only related to QSE, but also related to the allocated bandwidth. Therefore, we can jointly optimize QSE and allocated bandwidth to achieve the maximum QEE. In Fig. 6.4, with a fixed W, QEE is a bell-shaped curve function of QSE. With a fixed W, the increase of PSNR incurs the increase of QSE. However, a PSNR gain comes at the cost of a high transmission power. In the low QSE region, PSNR is low; so is the transmission power, making circuit power dominant in the total power consumption. Thus, the increase of QSE actually leads to a higher QEE, since a roughly flat power consumption is achieved, when transmission power is low. In the high QSE region, PSNR is also high, making the transmission power dominant in the total power consumption. As a result, the increase of QSE actually decreases QEE, which explains the bell-shaped curve of the QSE and QEE trade-off function.

Fig. 6.5 illustrates the impact of the decaying factor on QSE-QEE trade-off. Given the same bandwidth consumption (power consumption) and the same PSNR value, a higher decaying factor leads to a lower QSE (QEE). Thus a larger decaying factor tends to be more stingy on bandwidth or power consumption. Based on this observation, it is not difficult to understand why the QSE-QEE trade-off curve in Fig. 6.5 shifts towards right side when the bandwidth decaying factor θ decreases. When θ decreases and β is fixed, the cost of bandwidth consumption decreases. Thus the system can get a relatively higher

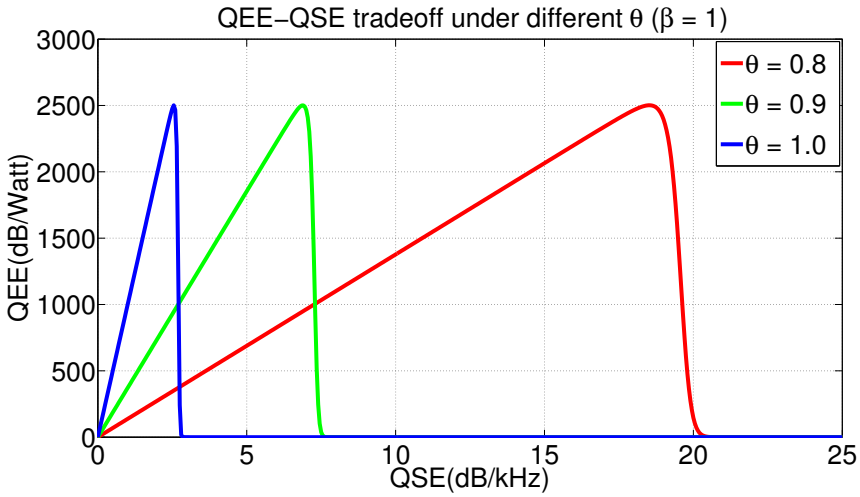

Fig. 6.5: QSE-QEE trade-off at different decaying factors θ, $\beta = 1$

PSNR on the cost of high bandwidth consumption. On the other hand, when θ increases and β is fixed, the cost of bandwidth consumption goes up and the system provides a relatively lower PSNR to save bandwidth. The same interpretation can be applied to β and energy-PSNR relationship. As such, QEE peaks at a relatively large QSE and at a high PSNR when θ is low. When θ value increases, QEE peaks at a smaller QSE and at a lower PSNR.

For comparison purpose, it is noted that both Fig. 6.5 and Fig. 1 of [45] have bell-shaped curves when considering circuit power. Furthermore, Fig. 6.6 depicts the trade-off between QEE and QSE when setting PSNR $= R$, $P_c = 0$, $\theta = 1$ and $\beta = 1$. QEE is a monotonously decreasing function of QSE in this case, which is consistent with the observations in Fig. 1 of [45].

6.4 QSE and QEE in PtP Rayleigh Fading Channel

In addition, we also consider a PtP Rayleigh fading channel scenario, in which we conduct the preliminary study of the system with analysis under fading conditions. Then, we formulate the data rate as

$$R = W\mathbb{E}\left\{\log_2\left(1 + \frac{P_t G}{\sigma^2}\right)\right\}, \tag{6.9}$$

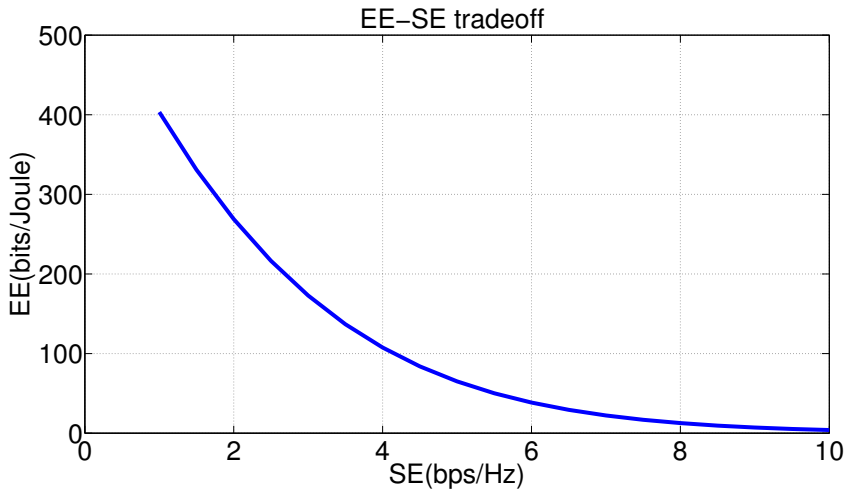

Fig. 6.6: EE-SE trade-off

where W is the allocated bandwidth and σ^2 is the variance of the AWGN. $G = |h|^2$ is the instantaneous channel power gain and obeys an exponential distribution with a probability density function $f(x) = \frac{1}{\Omega}e^{-\frac{x}{\Omega}}$, with Ω denoting the average channel power gain. Then QSE can be evaluated as:

$$\text{QSE} = \frac{\alpha \log_{10}\left[W\mathbb{E}\left\{\log_2\left(1+\frac{P_t G}{\sigma^2}\right)\right\}\right]}{W^{\theta}}. \tag{6.10}$$

It is noted that (6.10) is a monotonically increasing function of P_t. Then the total power consumption P is given by

$$P = P_t + P_c = \frac{P_t}{\zeta} + p + \xi\phi(W). \tag{6.11}$$

Here, ζ represents the drain efficiency of the power amplifier for transmission power P_t. Circuit power is modeled as $p + \xi\phi(W)$, where $p > 0$ represents a constant circuit power consumption component and $\xi > 0$ is the scale factor of the bandwidth-dependent function $\phi(W)$ [47]. Then we can express QEE as

$$\text{QEE} = \frac{\alpha \log_{10}\left[W\mathbb{E}\left\{\log_2\left(1+\frac{P_t G}{\sigma^2}\right)\right\}\right]}{\left(\frac{P_t}{\zeta} + p + \xi\phi(W)\right)^{\beta}}. \tag{6.12}$$

For the sake of representation simplicity, we denote SE as ψ and we have

$$\psi = \mathbb{E}\left\{\log_2\left(1+\frac{P_t G}{\sigma^2}\right)\right\} = \int_0^\infty \log_2\left(1+\frac{P_t x}{\sigma^2}\right)\frac{1}{\Omega}e^{-\frac{x}{\Omega}}dx = \log_2 e\cdot\left(e^{\frac{\sigma^2}{P_t\Omega}}\right)\cdot\Gamma\left(0,\frac{\sigma^2}{P_t\Omega}\right), \tag{6.13}$$

where $\Gamma(a,b)=\int_b^\infty \frac{1}{x^{1-a}e^x}dx$ is the upper incomplete gamma function [48].

Furthermore, we can establish the relationship between QSE and QEE as:

$$\text{QEE} = \frac{\alpha\log_{10}\left[W\mathbb{E}\left\{\log_2\left(1+\frac{P_t G}{\sigma^2}\right)\right\}\right]}{\left(\frac{P_t}{\zeta}+p+\xi\phi(W)\right)^\beta} = \frac{\text{QSE}\cdot W^\theta}{\left(\frac{P_t}{\zeta}+p+\xi\phi(W)\right)^\beta}. \tag{6.14}$$

Here P_t is denoted as the minimum transmission power required to achieve QSE and is given by

$$\min_{P_t>0} P_t \tag{6.15}$$

subject to

$$\frac{\alpha\log_{10}\left[W\log_2 e\cdot\left(e^{\frac{\sigma^2}{P_t\Omega}}\right)\cdot\Gamma\left(0,\frac{\sigma^2}{P_t\Omega}\right)\right]}{W^\theta} \geq \text{QSE}$$

To illustrate the fundamental relationship between QSE and QEE, we set $\alpha = 15$, $p = 100$mW, $\sigma^2 = -100$dBm, $W = 10$kHz, $\zeta = 0.4$, $\xi = 0.9$mW, and $\phi(W) = W/N_{ref}$, where $N_{ref} = 1$kHz is denoted as the reference bandwidth. Moreover, decaying factors θ and β increase from 0.6 to 1.0, with step size of 0.1. The average channel power gain $\Omega = |h|^2$ is evaluated to

$$\Omega = \Omega_0 d^{-4} \tag{6.16}$$

where $\Omega_0 = -70$dB is chosen as in [49], and d is the transmitter-receiver distance with value of 100m. For a given QSE, we can solve the transmit power P_t from (6.15) through simple line search algorithm due to its monotonicity.

In Fig. 6.7, with a fixed W, QEE is a bell-shaped curve function of QSE. This is because QSE is a monotonically increasing function of transmission power P_t. When P_t increases, the increase of PSNR incurs the increase of QSE. However, a PSNR gain comes at the cost of a high transmission power. In the low QSE region, PSNR is low; so is the transmission

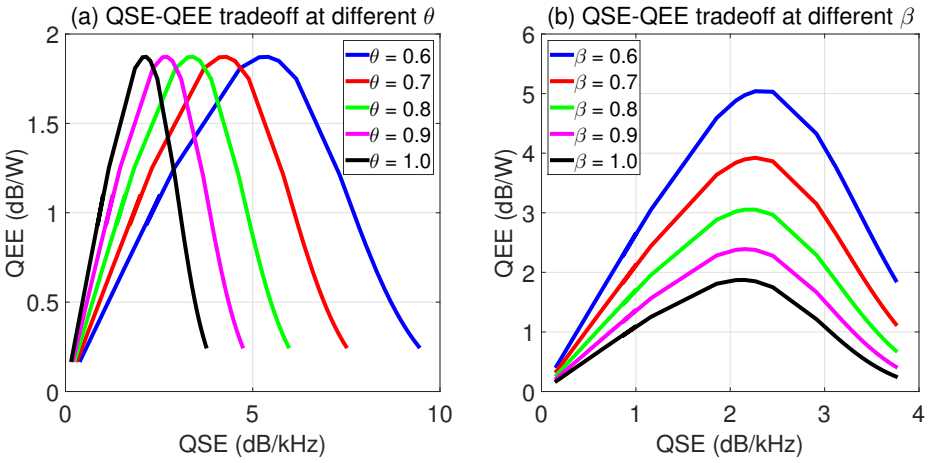

Fig. 6.7: QSE-QEE trade-off in Rayleigh fading channel

power, making circuit power dominant in the total power consumption. Thus, the increment of QSE actually leads to a higher QEE, since a roughly flat power consumption is achieved, when transmission power is low. In the high QSE region, PSNR is also high, making the transmission power dominant in the total power consumption. As a result, the increment of QSE actually decreases QEE, which explains the bell-shaped curve of the QSE and QEE trade-off function.

Fig. 6.7 also illustrates the impact of the decaying factor on QSE-QEE trade-off. Given the same bandwidth consumption (power consumption), a higher decaying factor leads to a lower QSE (QEE). Thus a larger decaying factor tends to be more stingy on bandwidth or power consumption. Based on this observation, it is not difficult to understand why the QSE-QEE trade-off curve in Fig. 6.7(a) shifts towards right side when the bandwidth decaying factor θ decreases. When θ decreases and β is fixed, the price of bandwidth consumption decreases. Thus the system can get a relatively higher PSNR on the price of high bandwidth consumption and achieve a relatively high QSE. On the other hand, when θ increases and β is fixed, the price of bandwidth consumption goes up and the system provides a relatively lower PSNR to save bandwidth, and the system achieves a relatively low QSE accordingly. As such, QEE peaks at a relatively large QSE and at a high PSNR when θ is low. When θ value increases, QEE peaks at a smaller QSE and at a lower PSNR. The same interpretation can be applied to Fig. 6.7(b).

6.5 QSE and QEE at System Level

We have studied the fundamental QSE and QEE performance in PtP AWGN and Rayleigh fading channels. In this section, we extend our QSE and QEE study to the system level. We will apply the proposed QSE and QEE to formulate a joint mobile association and resource allocation problem in a wireless heterogeneous network. A critical step in achieving efficient resource management is to associate UEs with proper serving BSs in order to fully exploit the network capacity/coverage/energy efficiency across different cell types. Most, if not all, of the existing mobile association studies in wireless heterogeneous networks are based on traditional spectrum and energy efficiency performance metrics [50–53], whereby video quality is transparent to spectrum and energy efficiencies during the mobile association process. This paper proposes a video quality-aware mobile association process to efficiently exploit capacity/coverage/energy/video quality gains, in a comprehensive manner.

Binary variable, $x_{i,j,k},\ i = 1, ..., N_c,\ j = 0, ..., N_r,\ k = 1, ..., N_u$, is defined as UE's association status, is expressed as follows:

$$x_{i,j,k} = \begin{cases} 1; & \text{if UE } k \text{ is associated with pBS } j \text{ in macro-cell } i,\ j = 0 \text{ when} \\ & \text{only associated with mBS } i \\ 0; & \text{otherwise.} \end{cases}$$

We first construct an $(N_c + N_p) \times N_u$ binary variable matrix $\mathbf{X}$ to indicate the association status for all the UEs:

$$\mathbf{X} \triangleq \begin{bmatrix} x_{1,0,1} & \cdots & x_{1,0,N_u} \\ \vdots & \vdots & \vdots \\ x_{N_c,N_r,1} & \cdots & x_{N_c,N_r,N_u} \end{bmatrix}_{(N_c+N_p)\times N_u}.$$

Next, the variable $n_{i,j,k}, i = 1, ..., N_c, j = 0, ..., N_r, k = 1, ..., N_u$ is defined to represent the bandwidth allocated to each UE. Consequently, we have:

$n_{i,0,k}$: bandwidth assigned to UE k if associated with mBS i;

$n_{i,j,k}$: bandwidth assigned to UE k if associated with pBS j in macro-cell i.

The bandwidth consumption matrices $\mathbf{N}_c$ and $\mathbf{N}_r$ for mBSs and pBSs are formulated as

$$\mathbf{N}_c \triangleq \begin{bmatrix} n_{1,0,1} & \dots & n_{1,0,N_u} \\ \vdots & \vdots & \vdots \\ n_{N_c,0,1} & \dots & n_{N_c,0,N_u} \\ 0 & \dots & 0 \\ \vdots & \vdots & \vdots \\ 0 & \dots & 0 \end{bmatrix}_{(N_c+N_p)\times N_u},$$

$$\mathbf{N}_r \triangleq \begin{bmatrix} 0 & \dots & 0 \\ \vdots & \vdots & \vdots \\ 0 & \dots & 0 \\ n_{1,1,1} & \dots & n_{1,1,N_u} \\ \vdots & \vdots & \vdots \\ n_{N_c,N_r,1} & \dots & n_{N_c,N_r,N_u} \end{bmatrix}_{(N_c+N_p)\times N_u}.$$

In a video supported heterogeneous network, mobile association and resource allocation aim to optimize video quality-aware SE and EE, i.e., QSE and QEE. Thus the decision consists of two parts for each UE: what BS to associate to (i.e., value of $x_{i,j,k}$), and how many resource are reserved for the video application at the association stage (i.e., value of $n_{i,j,k}$). We can formulate the system level mobile association and resource allocation problem as a MOOP, which is given by:

$$\mathbf{P1}: \max_{\mathbf{n},\mathbf{x}} \mathrm{QEE}(\mathbf{n},\mathbf{x}) \text{ and } \max_{\mathbf{n},\mathbf{x}} \mathrm{QSE}(\mathbf{n},\mathbf{x}) \tag{6.17}$$

subject to

$$
\begin{aligned}
Diag\left\{\mathbf{N}_c \times \mathbf{X}^T\right\}_{1:N_c} &\leq \mathbf{C}_m, && (6.18)\\
Diag\left\{\mathbf{N}_r \times \mathbf{X}^T\right\}_{N_c+1:N_c+N_p} &\leq \mathbf{C}_p, && (6.19)\\
Tr\left\{\mathbf{M}_k^T \times \mathbf{X}\right\} &\leq 1 \quad \text{for} \quad k = 1, \cdots, N_u,\\
\mathbf{N}_c + \mathbf{N}_r &\succ \mathbf{0}. && (6.20)
\end{aligned}
$$

Here, the system-wise QEE is defined as the ratio of system overall PSNR$(\mathbf{n}, \mathbf{x})$ and power consumption $P(\mathbf{n}, \mathbf{x})^{\beta}$. Similarly, QSE is defined as the ratio of system overall PSNR$(\mathbf{n}, \mathbf{x})$ and bandwidth consumption $W(\mathbf{n}, \mathbf{x})^{\theta}$, where $\mathbf{n}$ and $\mathbf{x}$ represent the vectors of $n_{i,j,k}$ and $x_{i,j,k}$, respectively. $\mathbf{C}_c \in \mathbb{R}^{+N_c \times 1}$ and $\mathbf{C}_r \in \mathbb{R}^{+N_p \times 1}$ are the total radio resources at mBSs and pBSs, respectively. (6.18) and (6.19) are the total bandwidth constraints at the mBSs and pBSs, respectively. (6.20) ensures one UE can at most associate with one mBS or one pBS. (6.20) ensures that the amount of allocated bandwidth is positive value.

The problem can be expressed in the following format equivalently:

$$
\min_{\mathbf{n},\mathbf{x}} \frac{1}{\text{QEE}(\mathbf{n}, \mathbf{x})} \text{ and } \min_{\mathbf{n},\mathbf{x}} \frac{1}{\text{QSE}(\mathbf{n}, \mathbf{x})} \tag{6.21}
$$

subject to (6.18)-(6.20).

We can solve the **P1** by applying weighted sum method [54], based on which the transformed problem is expressed as:

$$
\mathbf{P2}: \min U(\mathbf{n}, \mathbf{x}) = \frac{\omega_1}{\text{QEE}(\mathbf{n}, \mathbf{x})} + \frac{\omega_2}{\text{QSE}(\mathbf{n}, \mathbf{x})} \tag{6.22}
$$

subject to (6.18)-(6.20). And $\omega_m \in [0, 1]$ is the weight of the mth objective function which indicates the relative importance of that objective. Without loss of generality, it is a usual practice to choose weights such that the sum is one, i.e., $\sum_{m=1}^{2} \omega_m = 1$.

We can rewrite (6.22) as:

$$\mathbf{P3}: \min U(\mathbf{n},\mathbf{x}) = \frac{\omega_1 P(\mathbf{n},\mathbf{x})^{\beta} + \omega_2 W(\mathbf{n},\mathbf{x})^{\theta}}{\mathrm{PSNR}(\mathbf{n},\mathbf{x})}, \tag{6.23}$$

where $W(\mathbf{n},\mathbf{x})^{\theta}$, $P(\mathbf{n},\mathbf{x})^{\beta}$, and $\mathrm{PSNR}(\mathbf{n},\mathbf{x})$ are defined in (6.24)-(6.26).

$$\begin{aligned} W(\mathbf{n},\mathbf{x})^{\theta} &= \left[\rho_m Tr\left\{\mathbf{X}^T \times \mathbf{N}_c\right\} + \rho_p Tr\left\{\mathbf{X}^T \times \mathbf{N}_r\right\}\right]^{\theta} && (6.24)\\ P(\mathbf{n},\mathbf{x})^{\beta} &= (P_c + P_t)^{\beta} \\ &= Tr\left\{\left(\mathbf{P}_s + \xi\mathbf{X}^T \times \frac{\mathbf{N}_c + \mathbf{N}_r}{\mathbf{N}_{ref}}\right) + \mathbf{X}^T \times \left[\frac{\mathbf{P}_t \circ (\mathbf{N}_c + \mathbf{N}_r)}{\zeta}\right]\right\}^{\beta} && (6.25)\\ \mathrm{PSNR}(\mathbf{n},\mathbf{x}) &= Tr\left\{\left(\alpha \log_{10}\left[\mathbf{X}^T \times (\mathbf{N}_c + \mathbf{N}_r) \circ \log_2(\mathbf{I} + \mathbf{\Gamma})\right]\right)\right\}. && (6.26) \end{aligned}$$

ρ_m and ρ_p specify the relative cost of the bandwidth consumption at mBS and that at pBS. To achieve system load balancing, $\rho_m > \rho_p$ encourages more UEs to associate with pBSs. In (6.24), the first component represent the bandwidth consumptions from mBSs, and the second one represents the consumptions from PBSs. Moreover, P_c and $P_t \in \{P_m, P_p\}$ both are $(N_c + N_p) \times N_u$ power matrices, where the former one indicates the circuit power consumption and the latter one indicates the transmission power consumption. $\mathbf{P}_s$ is the static part of the circuit power consumption, and $\mathbf{N}_{ref}$ is a reference bandwidth. The dynamic part of the circuit consumption, which scales with the reference bandwidth with a proportional factor ξ [47]. Note that we also consider the power loss cost by the power amplifier with a ζ drain efficiency, and count this into the total transmission power consumption. PSNR is defined as a logarithm function of data rate with parameter α [43].

$\mathbf{\Gamma}$ is an $(N_c + N_p) \times N_u$ SINR matrix:

$$\mathbf{\Gamma} \triangleq \begin{bmatrix} \gamma_{1,0,1} & \cdots & \gamma_{1,0,N_u} \\ \vdots & \vdots & \vdots \\ \gamma_{N_c,0,1} & \cdots & \gamma_{N_c,0,N_u} \\ \gamma_{1,1,1} & \cdots & \gamma_{1,1,N_u} \\ \vdots & \vdots & \vdots \\ \gamma_{N_c,N_r,1} & \cdots & \gamma_{N_c,N_r,N_u} \end{bmatrix}_{(N_c+N_p)\times N_u},$$

of which entity $\gamma_{i,j,k}$ is defined as

$$\gamma_{i,0,k} = \frac{P_m h_{i,0,k}}{N_0 + \sum\limits_{i'=1,i'\neq i}^{N_c} h_{i',0,k} P_m + \sum\limits_{i'=1}^{N_c}\sum\limits_{j'=1}^{N_r} h_{i',j',k} P_p}, \tag{6.27}$$

$$\gamma_{i,j,k} = \frac{P_p h_{i,j,k}}{N_0 + \sum\limits_{\substack{i'=1 \\ (i',j')}}^{N_c} \sum\limits_{\substack{j'=1 \\ \neq(i,j)}}^{N_r} h_{i',j',k} P_p + \sum\limits_{i'=1}^{N_c} h_{i',0,k} P_m}, \tag{6.28}$$

where $h_{i,0,k}$ and $h_{i,j,k}$ represent the large scale channel gains between mBS i and UE k, between pBS j in macro-cell i and UE k, respectively.

$\mathbf{M}_k$ is an $(N_c + N_p) \times N_u$ all-zero matrix except the kth column, which is an all-one vector and given by

$$\mathbf{M}_k \triangleq [\mathbf{0} \cdots \mathbf{1_k} \cdots \mathbf{0}]_{(N_c+N_p)\times N_u}.$$

6.6 Nonlinear Fractional Programming

P3 is a mixed integer nonlinear non-convex combinatorial optimization problem. Non-linearity and non-convexity come from the fractional formulation of the objective function and the combinatorial nature comes from the binary association decision variables. It is difficult to solve the problem due to its high computational complexity, especially when considering a large number of decision variables, as in this case. Therefore, we solve the optimization problem in two tiers. In the outer tier, we apply the nonlinear fractional

programming to transform the optimization problem into a computational friendly form and search the optimal solution iteratively by using Dinkelbach's method [41]. In the inner tier, we use Lagrange dual decomposition to decompose the transformed problem into a sequence of similar sub-problems to search the optimal solution for each iteration point from the outer tier.

To facilitate the algorithm derivation, we express the objective function as:

$$\mathbf{P3}: \min U(\mathbf{n},\mathbf{x}) = \frac{\omega_1 P(\mathbf{n},\mathbf{x})^{\beta} + \omega_2 W(\mathbf{n},\mathbf{x})^{\theta}}{\mathrm{PSNR}(\mathbf{n},\mathbf{x})} = \frac{Q(\mathbf{n},\mathbf{x})}{D(\mathbf{n},\mathbf{x})}. \tag{6.29}$$

Without loss of generality, we define $S_{\mathbf{n}}$ as the feasible solution set of $\mathbf{n}$ and $S_{\mathbf{x}}$ as the feasible solution set of $\mathbf{x}$. Then **P3** is given by

$$q^* = \frac{Q(\mathbf{n}^*,\mathbf{x}^*)}{D(\mathbf{n}^*,\mathbf{x}^*)} = \min_{\mathbf{n}\in S_{\mathbf{n}},\mathbf{x}\in S_{\mathbf{x}}} \frac{Q(\mathbf{n},\mathbf{x})}{D(\mathbf{n},\mathbf{x})}. \tag{6.30}$$

Here, $\mathbf{n}^*, \mathbf{x}^*$ are the optimal solutions of $\mathbf{n}$ and $\mathbf{x}$, respectively.

Now we can state the following theorem.

***Theorem 1*:** q^* can be obtained if and only if

$$\mathbf{P4}: F(q^*) = \min_{\mathbf{n}\in S_{\mathbf{n}},\mathbf{x}\in S_{\mathbf{x}}} Q(\mathbf{n},\mathbf{x}) - q^* D(\mathbf{n},\mathbf{x}) = Q(\mathbf{n}^*,\mathbf{x}^*) - q^* D(\mathbf{n}^*,\mathbf{x}^*) = 0 \tag{6.31}$$

for any $Q(\mathbf{n},\mathbf{x}) > 0$ and $D(\mathbf{n},\mathbf{x}) > 0$.

Proof. Please refer to Appendix A. □

Up to this point, we transform the original optimization problem from a fractional form to a subtractive form. A closer look reveals that **P3** is equivalent to **P4**, which is computationally more tractable. To obtain the optimal solution of q, we use an iterative method, known as Dinkelbach's method [41], to solve **P4**. The method, summarized in Algorithm 3, guarantees convergence to the optimal solution (See Appendix B for proof). Notice that in step-7 of Algorithm 3, there exists an inner-tier optimization problem which

aims to solve the optimal solutions of $\mathbf{n}$ and $\mathbf{x}$ for a fixed q. We will discuss the inner-tier optimization in the following section.

Since $F(q^*) = 0$ is a very stringent convergence criteria, from a computational perspective, we use a very small positive value ϵ to denote convergence, i.e., $F(q^*) \leq \epsilon$ indicates convergence.

Algorithm 3 Outer-tier Iterative Nonlinear Fractional Programming

1: Initialize q, $\epsilon > 0$ as the convergence criteria, $i = 0$ is the iteration index and Max_i as the maximum iteration number.
2: Convergence = false.
3: **while** Convergence = false & $i < \text{Max}_i$ **do**
4: Update iteration index $i = i + 1$
5: Update $q = \frac{Q(\mathbf{n},\mathbf{x})}{D(\mathbf{n},\mathbf{x})}$
6: Insert q back to $F(q)$
7: Solve $F(q) = \min_{\mathbf{n}\in S_{\mathbf{n}},\mathbf{x}\in S_{\mathbf{x}}} Q(\mathbf{n},\mathbf{x}) - qD(\mathbf{n},\mathbf{x})$ (Detailed algorithm in Section 6.7)
8: **if** $F(q) < \epsilon$ **then**
9: Convergence = true.
10: $\mathbf{n}^* = \mathbf{n}, \mathbf{x}^* = \mathbf{x}$ and $q^* = q$
11: **end if**
12: **end while**
13: Output $\mathbf{n}^*, \mathbf{x}^*$ and q^*

6.7 Lagrange Dual Decomposition

Algorithm 3 includes an inner-tier optimization problem which aims to jointly solve the optimal resource allocations $\mathbf{n}$ and mobile association decisions $\mathbf{x}$ with a given q. The joint optimization problem is formulated as follows:

$$\mathbf{P5}: \min_{\mathbf{n}\in S_{\mathbf{n}},\mathbf{x}\in S_{\mathbf{x}}} Q(\mathbf{n},\mathbf{x}) - qD(\mathbf{n},\mathbf{x}) \tag{6.32}$$

subject to (6.18)-(6.20).

For a given q, **P5** is classified as a mixed integer nonlinear optimization problem that combines the combinatorial difficulty of optimizing discrete association variables with the challenges of optimizing resource allocation. It is an NP-hard problem. Although there exist traditional approaches such as brute-force and branch-and-bound methods to search

the global optimal solutions, it is nearly infeasible to solve it in real time for a large scale system. Therefore, we further introduce Lagrange dual decomposition to solve the complex optimization problem, leading to a reasonable computational complexity.

First we relax the integer variable $x_{i,j,k}$ to a real one in $[0,1]$, which can be interpreted as the association probability. And we introduce an auxiliary variable $\hat{n}_{i,j,k} = n_{i,j,k} \times x_{i,j,k}$. $\hat{n}_{i,j,k}$ can be considered as the actual radio resource allocation. The transformed optimization problem of **P5** is expressed as

$$\begin{aligned}
\mathbf{P6}: \min_{\hat{\mathbf{n}}} Q(\hat{\mathbf{n}}) - qD(\hat{\mathbf{n}}) &= \left[\omega_1 P(\hat{\mathbf{n}})^{\beta} + \omega_2 W(\hat{\mathbf{n}})^{\theta}\right] - q \times \mathrm{PSNR}(\hat{\mathbf{n}}) \\
&= \omega_1 Tr\left\{\mathbf{P}_s + \xi \times \frac{\hat{\mathbf{N}}_c + \hat{\mathbf{N}}_r}{\mathbf{N}_{ref}} + \frac{\mathbf{P}_t \circ \left(\hat{\mathbf{N}}_c + \hat{\mathbf{N}}_r\right)}{\zeta}\right\}^{\beta} \\
&+ \omega_2 \left(\rho_m Tr\left\{\hat{\mathbf{N}}_c\right\} + \rho_p Tr\left\{\hat{\mathbf{N}}_r\right\}\right)^{\theta} \\
&- q \times Tr\left\{\alpha \log_{10}\left[(\hat{\mathbf{N}}_c + \hat{\mathbf{N}}_r) \circ \log_2(\mathbf{I} + \mathbf{\Gamma})\right]\right\} \qquad (6.33)
\end{aligned}$$

subject to

$$Diag\left\{\hat{\mathbf{N}}_c \times \mathbf{I}^{\mathbf{T}}\right\}_{1:N_c} \leq \mathbf{C}_m, \qquad (6.34)$$

$$Diag\left\{\hat{\mathbf{N}}_r \times \mathbf{I}^{\mathbf{T}}\right\}_{N_c+1:N_c+N_p} \leq \mathbf{C}_p, \qquad (6.35)$$

$$Tr\left\{\mathbf{M}_k^T \times \mathbf{X}\right\} \leq 1 \quad \text{for} \quad k = 1, \cdots, N_u, \qquad (6.36)$$

$$\hat{\mathbf{N}}_c + \hat{\mathbf{N}}_r \geq \mathbf{0}. \qquad (6.37)$$

Although the relaxation only approximates the optimality of the original problem, it reduces the computational complexity greatly. **P6** can be proved to be quasi-convex with respect to $\hat{n}_{i,j,k}$ (see Appendix C). Furthermore, for the sake of reducing computational complexity, we assume $\beta = \theta = 1$. Then it is easy to prove that the relaxed optimization problem is strictly convex with respect to $\hat{n}_{i,j,k}$, and the constraints are linear functions. Based on Slator's condition [21], strong duality holds. Thus, we can obtain the primal solutions by solving the corresponding dual problem. For simplicity, we denote $\boldsymbol{\lambda}, \boldsymbol{\mu}, \boldsymbol{\nu}$

as the vectors for dual variables $\lambda_i > 0$, $\mu_{i,j} > 0$, and $\nu_k > 0$, respectively. Then the Lagrangian of **P6** is given by

$$\begin{aligned}
\mathcal{L}(\hat{\mathbf{n}}, \boldsymbol{\lambda}, \boldsymbol{\mu}, \boldsymbol{\nu}) &= \omega_1 Tr\left\{\mathbf{P}_s + \xi \times \frac{\hat{\mathbf{N}}_c + \hat{\mathbf{N}}_r}{\mathbf{N}_{ref}} + \frac{\mathbf{P}_t \circ \left(\hat{\mathbf{N}}_c + \hat{\mathbf{N}}_r\right)}{\zeta}\right\} \\
&+ \omega_2 \left(\rho_m Tr\left\{\hat{\mathbf{N}}_c\right\} + \rho_p Tr\left\{\hat{\mathbf{N}}_r\right\}\right) \\
&- q \times Tr\left\{\alpha \log_{10}\left[(\hat{\mathbf{N}}_c + \hat{\mathbf{N}}_r) \circ \log_2(\mathbf{I} + \boldsymbol{\Gamma})\right]\right\} \\
&+ \boldsymbol{\lambda} \times \left(Diag\left\{\hat{\mathbf{N}}_c \times \mathbf{I}^T\right\}_{1:N_c} - \mathbf{C}_m\right) + \boldsymbol{\nu} \times \left(Diag\left\{\mathbf{X}^T \times \mathbf{I}\right\} - \mathbf{1}\right) \\
&+ \boldsymbol{\mu} \times \left(Diag\left\{\hat{\mathbf{N}}_r \times \mathbf{I}^T\right\}_{N_c+1:N_c+N_p} - \mathbf{C}_p\right). \qquad (6.38)
\end{aligned}$$

The corresponding dual problem is formulated as

$$\max_{\boldsymbol{\lambda},\boldsymbol{\mu},\boldsymbol{\nu}} \inf_{\hat{\mathbf{n}}} \mathcal{L}(\hat{\mathbf{n}}, \boldsymbol{\lambda}, \boldsymbol{\mu}, \boldsymbol{\nu}). \qquad (6.39)$$

By applying dual decomposition technique [55], the above dual problem is converted to a sequence of similar sub-problems. Specifically, it can be separated into two levels of optimizations: low-level sub-problem and high-level master dual problem [56].

6.7.1 Low-level Sub-problem

In our case, the low-level sub-problem is to solve $\hat{\mathbf{n}}$ for the given dual variables:

$$\min_{\hat{\mathbf{n}}} \mathcal{L}(\hat{\mathbf{n}}, \boldsymbol{\lambda}, \boldsymbol{\mu}, \boldsymbol{\nu}). \qquad (6.40)$$

With KKT conditions [21], we take the partial derivative of $\mathcal{L}(\hat{\mathbf{n}}, \boldsymbol{\lambda}, \boldsymbol{\mu}, \boldsymbol{\nu})$ with respect to $\hat{\mathbf{n}}$ and set the partial derivative equal to zero:

$$\frac{\partial \mathcal{L}(\hat{\mathbf{n}}, \boldsymbol{\lambda}, \boldsymbol{\mu}, \boldsymbol{\nu})}{\partial \hat{\mathbf{n}}} = 0. \qquad (6.41)$$

By solving (6.41), the optimal solutions of $\hat{n}_{i,0,k}$ and $\hat{n}_{i,j,k}$ can be obtained from (6.42) and (6.43):

$$\hat{n}^*_{i,0,k} = x_{i,0,k}n^*_{i,0,k} = \left[\frac{\alpha q}{\ln 10 \times \left(\omega_1\left(P_s + \frac{\xi}{N_{ref}} + \frac{P_m}{\zeta_m}\right) + \omega_2\rho_m + \lambda_i\right)}\right]^+ \tag{6.42}$$

$$\hat{n}^*_{i,j,k} = x_{i,j,k}n^*_{i,j,k} = \left[\frac{\alpha q}{\ln 10 \times \left(\omega_1\left(P_s + \frac{\xi}{N_{ref}} + \frac{P_p}{\zeta_p}\right) + \omega_2\rho_p + \mu_{i,j}\right)}\right]^+, \tag{6.43}$$

where $[x]^+ = \max\{0, x\}$. Then we put the computed $\hat{\mathbf{n}}^*$ back to $\mathcal{L}(\hat{\mathbf{n}}, \boldsymbol{\lambda}, \boldsymbol{\mu}, \boldsymbol{\nu})$. In order to get optimal mobile association variables $\mathbf{x}^*$, we assume $x_{i,j,k} = 1$ (thus $\hat{\mathbf{n}}^* = \mathbf{n}^*$) and calculate the value of $\mathcal{L}(\mathbf{n}^*, \mathbf{x}, \boldsymbol{\lambda}, \boldsymbol{\mu}, \boldsymbol{\nu})$ in (6.44) and (6.45):

$$\begin{aligned}\mathcal{L}|_{x_{i,0,k}=1} &= \omega_1\left(\frac{P_m n^*_{i,0,k}}{\zeta_m} + \frac{\xi n^*_{i,0,k}}{N_{ref}} + P_s\right) + \omega_2\rho_m n^*_{i,0,k} - q\alpha\log_{10}\left(n^*_{i,0,k}\log_2(1+\gamma_{i,0,k})\right) \\ &+ \lambda_i n^*_{i,0,k} + \nu_k \end{aligned} \tag{6.44}$$

$$\begin{aligned}\mathcal{L}|_{x_{i,j,k}=1} &= \omega_1\left(\frac{P_p n^*_{i,j,k}}{\zeta_p} + \frac{\xi n^*_{i,j,k}}{N_{ref}} + P_s\right) + \omega_2\rho_p n^*_{i,j,k} - q\alpha\log_{10}\left(n^*_{i,j,k}\log_2(1+\gamma_{i,j,k})\right) \\ &+ \mu_{i,j} n^*_{i,j,k} + \nu_k \end{aligned} \tag{6.45}$$

In order to satisfy the constraint that each UE can at most associate with one mBS or one pBS, the optimal mobile association decision for UE k is then given by:

$$x^*_{i,j,k} = \begin{cases} 1; & \text{for } \{i,j\} = \arg\min \mathcal{L}|_{x_{i,j,k}=1}, \forall i,j \\ 0; & \text{otherwise} \end{cases} \tag{6.46}$$

6.7.2 High-level Master Dual Problem

The high-level master dual problem is to update the dual variables $\boldsymbol{\lambda}$, $\boldsymbol{\mu}$, and $\boldsymbol{\nu}$. In our case, the dual function is given by:

$$\max_{\boldsymbol{\lambda},\boldsymbol{\mu},\boldsymbol{\nu}} g(\boldsymbol{\lambda}, \boldsymbol{\mu}, \boldsymbol{\nu}) = \mathcal{L}(\mathbf{n}^*, \mathbf{x}^*, \boldsymbol{\lambda}, \boldsymbol{\mu}, \boldsymbol{\nu}). \tag{6.47}$$

Since $g(\boldsymbol{\lambda}, \boldsymbol{\mu}, \boldsymbol{\nu})$ is differentiable, we can solve the master dual problem with a gradient descent method. The Lagrange multipliers can be updated by:

$$\lambda_i(t+1) = \left[\lambda_i(t) - \eta_1(t)\left(\mathbf{C}_m(i) - \sum_{k=1}^{N_u} x^*_{i,0,k} n^*_{i,0,k}\right)\right]^+, \tag{6.48}$$

$$\mu_{i,j}(t+1) = \left[\mu_{i,j}(t) - \eta_2(t)\left(\mathbf{C}_p(i,j) - \sum_{k=1}^{N_u} x^*_{i,j,k} n^*_{i,j,k}\right)\right]^+, \tag{6.49}$$

$$\nu_k(t+1) = \left[\nu_k(t) - \eta_3(t)\left(1 - \sum_{i=1}^{N_m}\sum_{j=0}^{N_p} x^*_{i,0,k}\right)\right]^+, \tag{6.50}$$

where t is the iteration index and $\{\eta_1(t), \eta_2(t), \eta_3(t)\}$ are sufficiently small positive step sizes. We apply diminishing step size rule by setting $\eta_{1,2,3}(t) = (1+s)/(t+s)$, where s is a non-negative number. By choosing diminishing step size, the iteration can converge to the optimal value for bounded gradients [55].

6.7.3 Iterations between Low-level and High-level

We can solve the dual problem by solving the low-level and high-level problems iteratively. We feed the updated multipliers' solutions from the high-level master dual problem to the low-level sub-problem. Then with the optimal solutions of resource allocation and mobile association in the low-level problem, we can update them back to the master problem and re-calculate the multipliers. When the iteration converges, the dual problem is considered solved. We can go back to Algorithm 3 to find out the optimal q^*.

For additional clarity, we summarize the two-tier optimization process in Fig. 6.8.

6.8 Complexity Analysis

Computational complexity needs to be addressed in implementing joint mobile association and resource allocation schemes for heterogeneous networks. The joint mobile association and resource allocation problem is commonly known as an NP-hard problem and an exhaustive search solution is computationally prohibitive. The proposed scheme aims to reduce the computational complexity by using Dinkelbach's method and dual de-

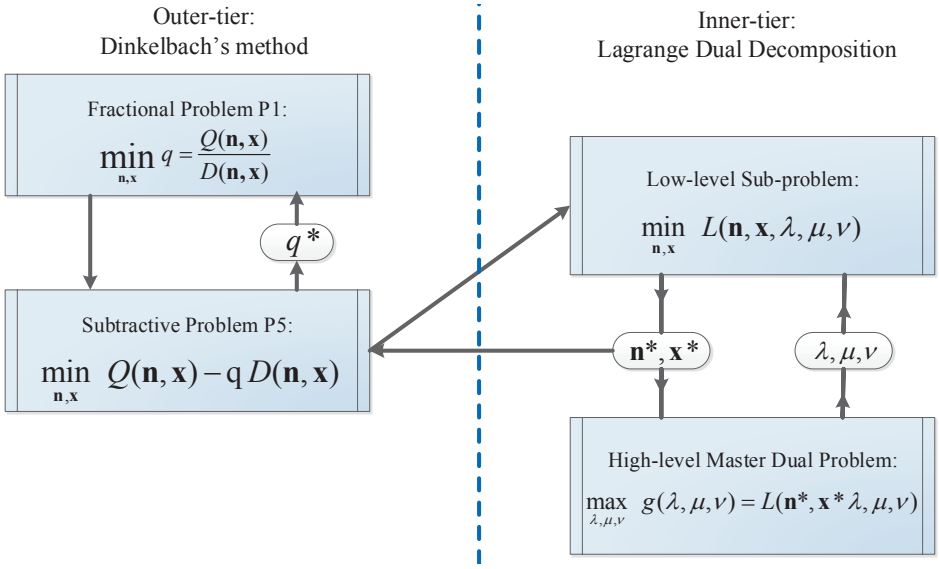

Fig. 6.8: Two-tier optimization process

composition approach. Let T_{out} represent the number of iterations required for outer-tier Dinkelbach's method to converge, and T_{in} represent the number of iterations required for the gradient descent method in (6.48)-(6.50) to converge. Both T_{out} and T_{in} are related to convergence criteria and are restricted by pre-set maximum iteration numbers. Starting from the inner-tier dual decomposition, the computational complexity is dominated by the calculation of $\hat{\mathbf{n}}$ and T_{in}. The computational complexity is $\mathcal{O}\left(2T_{in}N_u(N_m + M_p)\right)$, where $\mathcal{O}(\cdot)$ is the big-O notation. Thus, the overall algorithmic complexity with outer-tier Dinkelbach's method is $\mathcal{O}\left(2T_{out}T_{in}N_u(N_m + M_p)\right)$. It is observed that the proposed algorithm is linear in the number of UEs N_u. A brute-force search algorithm has a complexity of $\mathcal{O}\left(T_{out}(N_m + M_p)^{N_u}N^{N_u}\right)$, where N is the number of bandwidth assignment options for each MU. The overall algorithmic complexity is exponential in N_u. Therefore, the proposed scheme can achieve a major reduction in the computational complexity.

6.9 Performance Evaluation

In this section, we conduct simulations by following 3GPP specified cases in [22]. We consider a 19-cell 3-sector three-ring hexagonal network structure, where one mBS is located in the center of each macro-cell, and 4 pBSs are equally-distanced deployed in the overlaid

pico-cells within each macro-cell, which form a two-tier heterogeneous network. UEs are uniformly distributed over the network. Standard SVC test video sequence *Foreman* [57] in the QCIF format (176×144 pixels) is used in the simulation, and the content-related parameter $\alpha = \{10, 12\}$. For the sake of generality, half of the UEs download video frames with $\alpha = 10$, and the other half requires the video with $alpha = 12$. Additional system parameter settings are shown in Table 6.1.

Fig. 6.9 illustrates the trade-off between QSE and QEE for UEs with different video contents requirements. Specifically, we change the weight $0 < \omega_i < 1, \forall i \in \{1, 2\}$ uniformly with a step size of 0.1, such that $\omega_1 + \omega_2 = 1$. For each specific weight pair, we obtain the optimal mobile association and resource allocation solutions. The trade-off curve, which is also named Pareto front, is achieved by the considered mobile association and resource allocation. It is observed that QEE decreases with the increment of QSE. This is because with a higher weight on QSE, the system gives a higher priority to QSE maximization objective. Thus, more bandwidths are consumed to increase the system QSE. Whilst, more bandwidth consumption results in a higher transmit power consumption. Hence, the increment of QSE will cause the decrement of QEE. In addition, UEs with a higher video content parameter will lead to a higher QSE and QEE values. This is because a higher α will result in a higher PSNR value at the same data rate.

By setting ω_1 or ω_2 equal to zero, the MOOP turns into a single optimization problem for QSE and QEE, respectively. By adjusting the decaying factors, we evaluate the video qualities and utilization under QSE and QEE cases.

Fig. 6.10 illustrates the average PSNR values and mean option score (MOS) scales in the system at different decaying factors. Fig. 6.10(a1) and Fig. 6.10(b1) illustrate the average PSNR values in the system at different decaying factors. With a larger decaying factor, it is not cost-efficient to improve the QSE/QEE performance of the system by increasing the PSNR. This in turn explains why the average PSNR value is low at a high decaying factor. For a reduced decaying factor, the system improves its QSE/QEE by consuming more bandwidth and power, at an acceptable cost. Therefore, it can be observed that the

Table 6.1: System parameter settings

Parameter	**Settings**
mBS	57
pBS	4 per macro-cell
UE	300 per cell
Static Circuit Power	$P_s = 10$ mWatt
Proportional Factor	$\xi = 20$ mWatt
Drain Efficiency	$\zeta_m = 35\%, \zeta_p = 20\%$
Transmit Power	$P_m = 30$ Watt, $P_p = 1$ Watt
System/Reference Bandwidth	20 MHz
Noise Model and density	AWGN, -174 dBm/Hz
Path loss from MBS to UE	$PL_{LOS}(R) = 103.4 + 24.2\log_{10}(R)$ $PL_{NLOS}(R) = 131.1 + 42.8\log_{10}(R)$
Path loss from PBS to UE	$PL_{LOS}(R) = 103.8 + 20.9\log_{10}(R)$ $PL_{NLOS}(R) = 145.4 + 37.5\log_{10}(R)$
Shadowing	8 dB, log-normal std. deviation

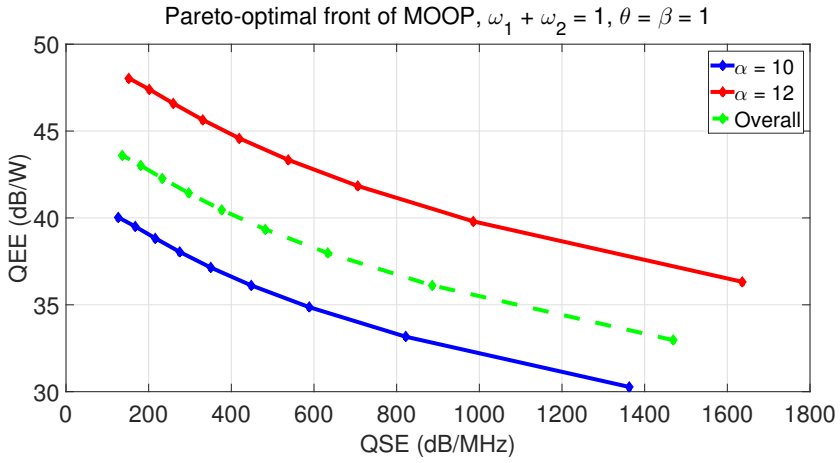

Fig. 6.9: Pareto-optimal front of MOOP, $\omega_1 + \omega_2 = 1, \theta = \beta = 1$

PSNR values are relatively higher when the decaying factor is lower. Furthermore, we map PSNR into MOS ITU 5-point scale [58] in Table 6.2, from which a subjective QoE, i.e., MOS scale, can be obtained from an objective QoE, i.e., PSNR. As shown in Fig. 6.10(a2)

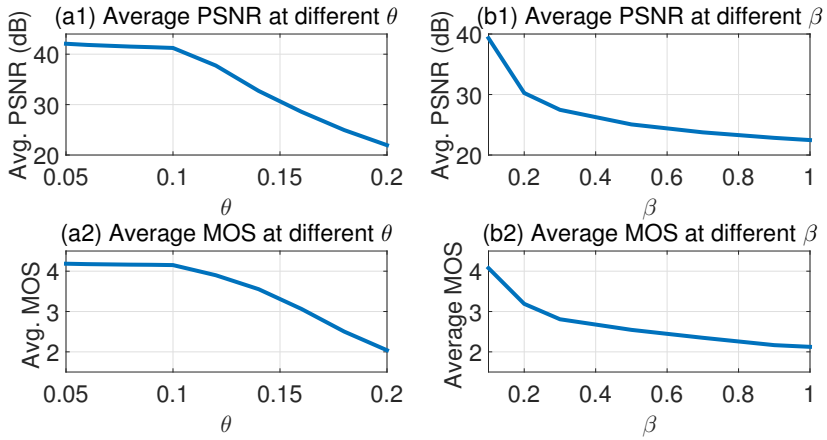

Fig. 6.10: Average PSNR and MOS at different decaying factors : (a) QSE-optimized; (b) QEE-optimized

and Fig. 6.10(b2), MOS scales and PSNR values follow similar trends, i.e., a higher PSNR value corresponds to a better MOS scale and vice versa.

Table 6.2: Possible PSNR to MOS conversion

PSNR (dB)	**MOS**
≥ 37	5 (Excellent)
31–37	4 (Good)
25–31	3 (Fair)
20–25	2 (Poor)
≤ 20	1 (Bad)

Fig. 6.11 depicts the PSNR distributions with different decaying factors. Considering Fig. 6.11(b), for example, it can be observed that the system with a lower power decaying factor has a better PSNR distribution. It is also shown that the curve of $\beta = 0.1$ has an approximate 15dB gain on the mean PSNR value over the curve of $\beta = 1.0$. With a smaller decaying factor, more UEs can receive a higher PSNR due to the relatively low cost of bandwidth/power consumption. When the decaying factor is large, the system is stingy with bandwidth/power consumption. Thus, UEs are discouraged to consume more bandwidth/power to increase the PSNR values.

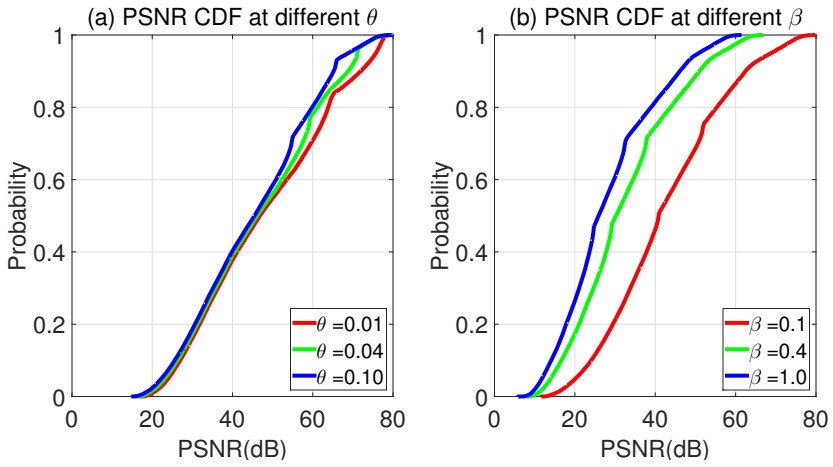

Fig. 6.11: PSNR CDF at different decaying factors : (a) QSE-optimized; (b) QEE-optimized

Fig. 6.12 explores the MOS distribution among UEs at different decaying factors. It is shown that the system with a lower decaying factor has more UEs achieving a better MOS scale. For example, when $\beta = 0.1$, almost 10000 of total 17100 UEs can experience an excellent perceived video quality of MOS scale 5. While when $\beta = 1.0$, more UEs suffer from poor and bad perceived video qualities with MOS scales 1 and 2, which might lead to a higher churn rate. This is because with a smaller decaying factor, more UEs can receive a higher PSNR (corresponds to a better MOS scale) due to the relatively low cost of bandwidth/power consumption. When the decaying factor is large, the system is stingy with bandwidth/power consumption. Thus, UEs are discouraged to consume more bandwidth/power to increase the PSNR values, thereby they obtain a worse MOS scale.

To further elaborate the difference between the traditional SE/EE and QSE/QEE defined in this paper, we run system level simulations by maximizing system SE and EE and compare the performance with the proposed QSE and QEE framework. To set up a fair comparison, in both SE/EE and QSE/QEE simulations, the number of UEs served and the maximum bandwidth/energy consumed in both cases are the same. But one aims to maximize overall SE/EE and one aims to maximize overall QSE/QEE. Fig. 6.13 depicts the UEs distribution with different MOS scales, in which it observed that more UEs obtain higher MOS scales by using QSE/QEE as performance objectives. This shows that QSE

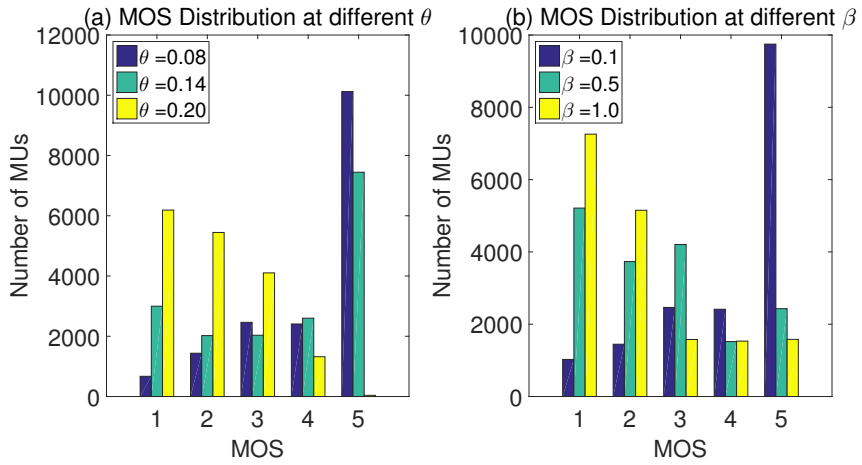

Fig. 6.12: MOS distribution at different decaying factors : (a) QSE-optimized; (b) QEE-optimized

and QEE utilize the given bandwidth/energy more intelligently so that a better system level video QoE is achieved while SE/EE objectives fail to capture QoE in its definition.

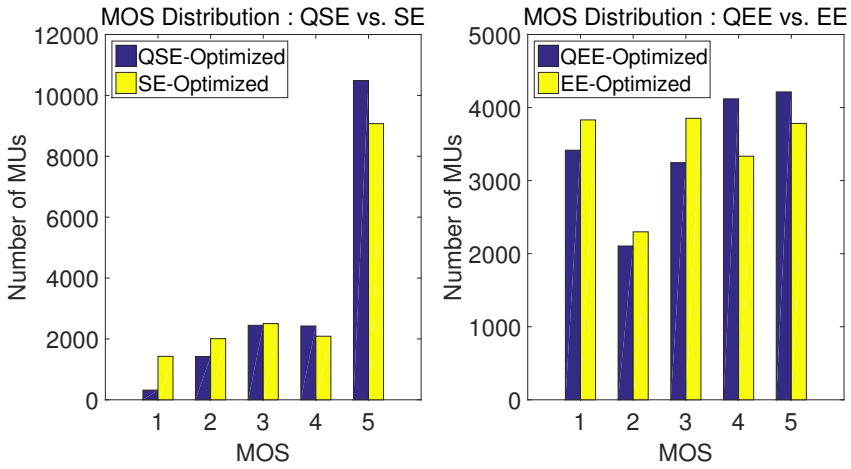

Fig. 6.13: Comparison of MOS

In Fig. 6.14, we analyze the impacts of weight factors ρ_m and ρ_p on mBSs' and pBSs' resource utilization. As the ρ_m/ρ_p increases, it is noted that the mBS's utilization decreases (Fig. 6.14(a)) while the pBS's utilization increases from 30% at $\rho_m/\rho_p = 2$ to 88% at $\rho_m/\rho_p = 8$ (Fig. 6.14(b)). At the same time, more UEs associate with pBSs (Fig. 6.14(d)) while the portion of the total UEs attached to mBSs drops from 40% at $\rho_m/\rho_p = 2$ to 27% at $\rho_m/\rho_p = 8$ (Fig. 6.14(c)). With a higher ρ_m/ρ_p ratio, the UE associated with an mBS

has a higher bandwidth consumption cost. A higher resource cost will divert more UEs to associate with pBSs which have relatively lower cost. Considering that pBS operates in a low power, the system can achieve a better energy-efficiency goal by using pBSs to achieve load balancing.

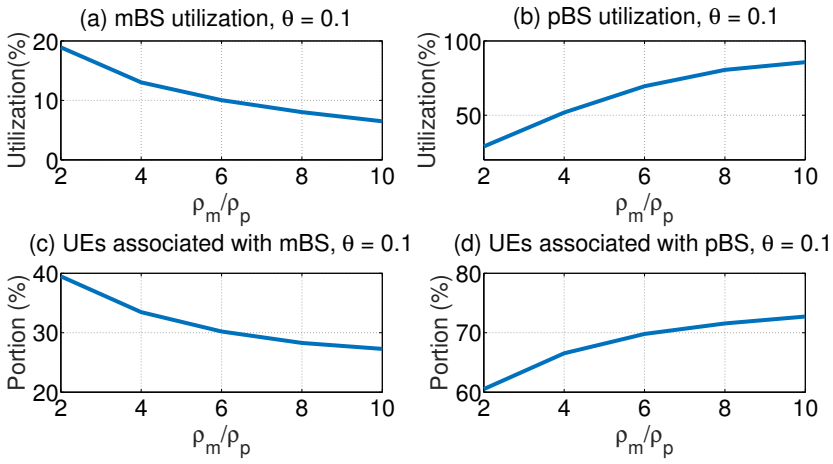

Fig. 6.14: Utilization of mBS and pBS at different ρ_m and ρ_p

6.10 Chapter Summary

In this chapter, we developed a video QSE and QEE based mobile association and resource allocation scheme in a wireless heterogeneous network. We first explored the fundamental trade-off between QSE and QEE in a PtP AWGN channel and in a Rayleigh fading channel, respectively. We then formulated a joint optimization problem of mobile association and resource allocation in a heterogeneous network. The formulation was also extended to explore the trade-off study between QSE and QEE at the system level. The optimization was solved by using nonlinear fractional programming and Lagrange dual decomposition in a computational efficient way. Simulation results show that by using QSE/QEE objectives the system can achieve a much better video quality than the traditional SE/EE objectives. Moreover, we consider both objective assessment (i.e., PSNR) and subjective assessment (i.e., MOS) to evaluate the reception video quality, and look into the connections between PSNR and MOS for a convincing QoE performance. The study further shows that the sys-

tem performance also greatly depends on the bandwidth and power decaying factors defined in QSE/QEE. The results pave the way for future research to gain better understanding of how the decaying factors relate to the bandwidth and energy pricing models in wireless heterogeneous networks from economy perspectives.

Chapter 7

Trade-offs in Video Transmission over Wireless Heterogeneous Networks: Energy, Bandwidth and QoE

7.1 Introduction

The recent surge of video traffic is stressing the mobile and wireless network infrastructure, pushing its capacity beyond its limit. Streaming video is gradually becoming an integral part of typical daily activities in different settings, ranging from home applications and Internet services to video collaboration and video conferencing in business and academic environments. It is anticipated that this trend will continue in the future at a faster rate, with video traffic exceeding 80% of consumer Internet traffic [59]. The exponential growth of video traffic will undoubtedly have a significant impact on the energy and bandwidth consumption of future wireless infrastructure, greatly challenging their ability to deliver the users' expected QoS and QoE. Addressing the stringent requirements of mobile video streaming is a daunting challenge that must be addressed in next generation wireless network infrastructure.

To address holistically the above challenge, we formulate a multi-objective optimization problem in this chapter to explore the trade-off relationships among energy consumption, bandwidth consumption and perceived video qualities. We aim to search for the Pareto optimal mobile association and resource allocation in a video transmitted wireless heterogeneous network. The multiple objectives focus on spectrum efficiency, energy efficiency and QoE. The weighted Tchebycheff approach is introduced to combine multiple objectives and formulate a mixed integer nonlinear programming (MINLP) problem [60]. In order to make this problem tractable, we apply a linear programming relaxation and variable transformation to reduce the computation complexity. The transformed problem is convex

and is solved by a sequence of sub-problems via dual decomposition technique [55].

7.2 Problem Formulation

Without loss of generality, we consider the same system model of the same as the previous chapter as shown in Fig. 5.1, where macro-cells and pico-cells coexist in the overlay mode in a downlink heterogeneous cellular network. In each macro-cell, one mBS is located in the center, overlaid with several uniformly deployed pBSs. mBS typically transmits at a high power level to provide blanket coverage and seamless mobility, while pBS transmits at substantially a lower power and aims to eliminate coverage holes, to improve the capacity in hot spots, as well as to provide traffic offloading.

To study the holistic system design problem, which has multiple possibly conflicting performance metrics, we formulate a multi-objective optimization problem (MOOP). The main objective is to maximize the users' perceived video quality (PSNR) and minimize energy consumption and bandwidth consumption. As defined before, N_c is denoted as the total number of mBSs, N_r is the number of pBSs and N_u is the number of UEs. The transmit power of mBS is P_m and the transmit power of pBS is P_p, where $P_m >> P_p$. The disparity of transmit power at different BSs results in different coverage ranges. To fully exploit the heterogeneous network capacity and coverage gains provided by multi-tier resources, associating UEs to the proper serving BSs is critical.

The following binary variables, $x_{i,j,k}$, $i = 1, \cdots, N_c$, $j = 0, \cdots, N_r$, $k = 1, \cdots, N_u$, are used to indicate UE's association status.

$$x_{i,0,k} = \begin{cases} 1; & \text{if UE } k \text{ is associated with mBS } i \\ 0; & \text{otherwise} \end{cases}$$

$$x_{i,j,k} = \begin{cases} 1; & \text{if UE } k \text{ is associated with pBS } j \text{ in macro-cell } i \\ 0; & \text{otherwise.} \end{cases}$$

Furthermore, variables $n_{i,j,k}, i = 1, \cdots, N_c, j = 0, \cdots, N_r, k = 1, \cdots, N_u$ are denoted

as the network resources allocated to each UE.

$n_{i,0,k}$: network resource assigned to UE k associated with mBS i;

$n_{i,j,k}$: network resource assigned to UE k associated with pBS j in macro-cell i.

We consider a video downloading service, where each UE sends a request to either an mBS or a pBS. Upon receiving the request, mBS or pBS will choose the archival video file from the video server and start downloading. In video transmission, different spatial resolution and frame rates result in different data rate requirements, which can be achieved through different video coding schemes. In this chapter, we use the term layered video and scalable video interchangeably. Based on the scalable video coding, a video frame is usually encoded into a base layer and multiple enhancement layers [61]. The enhancement layers can only be decoded when the base layer is received intactly. Enhancement layers can refine and improve the perceived video quality. The more the layers are received, the better the video quality is provided. The perceived video quality is normally measured by the PSNR, which can be approximated as a log-function of the received data rate [43]:

$$\text{PSNR} = \alpha \log_{10}(R) = \alpha \log_{10}(n \log_2(1+\gamma)), \tag{7.1}$$

where α is the content-related video parameter. For a video with many dynamic scenes, α value is high. R is denoted as the received data rate and γ represents the SINR for the video transmitted wireless channel.

The received SINR for UE k associated with mBS i is expressed as:

$$\gamma_{i,0,k} = \frac{P_m |h_{i,0,k}|^2}{\sum\limits_{i' \neq i}^{N_m} \sum\limits_{k=1}^{N_u} P_m |h_{i',0,k}|^2 \sum\limits_{i'=1}^{N_m} \sum\limits_{j'=1}^{N_r} \sum\limits_{k=1}^{N_u} P_p |h_{i',j',k}|^2 N_0}, \tag{7.2}$$

where $h_{i,0,k}$ and $h_{i,j,k}$ represent the channel gain between UE k and mBS i, and the channel gain between UE k and pBS j in ith macro-cell, respectively. N_0 denotes the thermal noise level. Similarly, the SINR for UE k associated with the pBS j in ith macro-cell is expressed

as:

$$\gamma_{i,j,k} = \frac{P_p|h_{i,j,k}|^2}{\sum_{i'=1}^{N_m}\sum_{k=1}^{N_u} P_m|h_{i',0,k}|^2 \sum_{i'\neq i}^{N_m}\sum_{j'\neq j}^{N_r}\sum_{k=1}^{N_u} P_p|h_{i',j',k}|^2 N_0}. \tag{7.3}$$

In the following, we start with the single objective optimization formulation, and discuss the possible optimization outcomes and the corresponding limitations of a single objective optimization. The results of this optimization will be used to motivate the multi-objective optimization that can lead to a better holistic system performance.

7.2.1 Objective 1: Perceived Video Quality Maximization

For the video downloading service in wireless heterogeneous networks, the first objective considered is to maximize the user's perceived video quality. The first optimization problem can be formulated as

$$\mathbf{P1}: \max \sum_{i=1}^{N_c}\sum_{j=0}^{N_r}\sum_{k=1}^{N_u} x_{i,j,k}\mathrm{PSNR}_{i,j,k} \tag{7.4}$$

$$\text{C1} \quad : \quad \sum_{i=1}^{N_c}\sum_{k=1}^{N_u} x_{i,0,k} n_{i,0,k} \leq C_i^m \tag{7.5}$$

$$\text{C2} \quad : \quad \sum_{i=1}^{N_c}\sum_{j=1}^{N_r}\sum_{k=1}^{N_u} x_{i,j,k} n_{i,j,k} \leq C_{i,j}^p \tag{7.6}$$

$$\text{C3} \quad : \quad \sum_{i=1}^{N_c}\sum_{j=0}^{N_r}\sum_{k=1}^{N_u} x_{i,j,k} \leq 1 \tag{7.7}$$

$$\text{C4} \quad : \quad \mathrm{PSNR}_{i,j,k} \geq \mathrm{PSNR}_{min} \tag{7.8}$$

$$\text{C5} \quad : \quad n_{i,j,k} \geq 0, \tag{7.9}$$

$$\text{C6} \quad : \quad x_{i,j,k} \in \{0,1\}. \tag{7.10}$$

Here, C1 and C2 are the resource constraints at the mBSs and pBSs, respectively. C3 guarantees that each UE can, at most, associate with one BS, whether mBS or pBS. C4 ensures that each UE receives the base video layer so that the video can be successfully decoded and replayed. C5 ensures that the radio resource allocation is non-negative. C6

ensures that the association variable is a binary value. In order to achieve the maximum system level PSNR, the network must first guarantee that all users accepted by the BSs will meet the minimum PSNR. For each user, attaching to the node that provides the best downlink SINR will be beneficial to PSNR. After allocating the minimum PSNR equivalent resources to all the UEs, each BS will then distribute the leftover resources to the single user that has the best channel condition with that BS in order to achieve the highest system sum PSNR, and thereby significantly impairing the fairness among users. Thus, even a single PSNR objective will not lead to any good design guideline in reality. However, by jointly considering other performance metrics, one may define a more practically meaningful problem, as shown later.

7.2.2 Objective 2: Energy Efficiency

The second objective considered is to minimize the total power consumption at BS, since the total BS power consumption in the system is dominant over the UE power consumption. An mBS and a pBS each has a different transmit power. By attaching UEs to the proper nodes, we can achieve system level energy efficiency. The power model we apply in this paper is given by

$$\begin{aligned} P_{i,0,k} &= P_c + P_m n_{i,0,k}, \quad \text{for UE } k \text{ associated with mBS } i \\ P_{i,j,k} &= P_c + P_p n_{i,j,k}, \quad \text{for UE } k \text{ associated with pBS } j. \end{aligned}$$

Here, P_c is denoted as the static circuit power consumption. The second optimization problem can be formulated as:

$$\mathbf{P2}: \min \sum_{i=1}^{N_c} \sum_{j=0}^{N_r} \sum_{k=1}^{N_u} x_{i,j,k} \left(P_t n_{i,j,k} + P_c \right), P_t \in \{P_m, P_p\} \tag{7.11}$$

subject to C1-C6.

Since mBS has a much higher transmit power than pBS, from an energy efficiency perspective, most UEs will be associated with their closest respective pBSs, unless certain

UEs get close enough to mBSs to cause the low bandwidth (due to good SINR) consumption to offset the high power consumption from mBSs. In this case, the coverage range of mBSs virtually shrinks, which defeats the purpose of using high power for blanket coverage and mobility, not to mention video QoE is not considered at all in this model. Either a more practical power model needs to be defined, or other performance metrics need to be considered in order to define a better system design.

7.2.3 Objective 3: Network Resource Efficiency

Video transmission over wireless networks is intensely bandwidth consuming. Therefore mobile association can alternatively minimize network resource consumptions. A wireless heterogeneous network consists of BSs from different tiers, each serving different coverage and capacity needs. In a traditional homogeneous network consisting mainly of mBSs, association based on downlink SINR strength will normally lead to the most system-wise resource efficiency if users are uniformly distributed and little load balancing is needed. In a heterogeneous network, even though users are uniformly distributed, due to different transmit powers of different BSs, range expansion is needed for pBSs to have enough coverage to ensure that their resources will be best utilized. In the mobile association that aims to achieve a high system-wise resource efficiency, we need to offload traffic to pBSs to effectively expand their coverage. In our model, we give different weights to different BSs. By adjusting weights, load balancing is achieved. The optimization model is formulated as

$$\mathbf{P3}: \min \sum_{i=1}^{N_c} \sum_{j=0}^{N_r} \sum_{k=1}^{N_u} x_{i,j,k} \left(\rho_m n_{i,0,k} + \rho_p n_{i,j,k} \right) \tag{7.12}$$

subject C1-C6.

Here, ρ_m and ρ_p specify the relative costs of the network resource consumptions at mBSs and at pBSs, respectively. For the purpose of system load balancing and energy efficiency, pBS can be given a lower cost than mBS, so that more UEs can be associated with pBSs.

7.2.4 Multi-objective Optimization Problem

Each of the above three objectives targets a different performance goal. These various objectives may not be able to reach the respective optimal solution under the same settings. Sometimes they could even conflict with each other. Hence, it is necessary to consider the system design problem from a multi-objective perspective by considering compromise and trade-off. In this part, we redefine a new MOOP formulation by considering the above three individual objectives altogether:

$$\mathbf{MOOP1}: \quad \min -\sum_{i=1}^{N_c}\sum_{j=0}^{N_r}\sum_{k=1}^{N_u} x_{i,j,k}\mathrm{PSNR}_{i,j,k} \tag{7.13}$$

$$\min \sum_{i=1}^{N_c}\sum_{j=0}^{N_r}\sum_{k=1}^{N_u} x_{i,j,k}\left(P_t n_{i,j,k} + P_c\right) \tag{7.14}$$

$$\min \sum_{i=1}^{N_c}\sum_{j=0}^{N_r}\sum_{k=1}^{N_u} x_{i,j,k}\left(\rho_m n_{i,0,k} + \rho_p n_{i,j,k}\right) \tag{7.15}$$

subject to C1-C6. Note that the objective function **P1** is transferred into an equivalent minimization form to facilitate the presentation.

7.3 Weighted Tchebycheff Approach and Dual Decomposition

Before solving the MOOP, we first present the following two definitions [60].

Definition 1: A point, $x^* \in \mathbf{X}$, is Pareto optimal iff there does not exist another point, $x \in \mathbf{X}$, such that $\mathbf{F}(x) \leq \mathbf{F}(x^*)$, and $F_i(x) < F_i(x^*)$ for at least one function F_i.

Definition 2: A point, $\mathbf{U}^\circ \in \mathbf{Z}$, is a utopia point iff for each $i = 1, 2 \cdots, k$, $U_i^\circ = \min_x \{U_i(\mathbf{x}) | \mathbf{x} \in \mathbf{X}\}$.

For the sake of simplicity, we denote the objective function for the yth objective as $U_y(\mathbf{x}, \mathbf{n})$, $y \in \{1, 2, 3\}$. The utopia point, which corresponds to the optimal value by optimizing the objective individually, is denoted as U_y°. Please refer to our previous work [42] for the details to obtain the utopia point (single objective optimization). [60] introduces a number of multi-objective optimization methods to obtain the Pareto optimal points. The set of Pareto optimal points is called Pareto frontier or Pareto optimal set. We ap-

ply weighted Tchebycheff approach and dual decomposition to obtain the set of all Pareto optimal mobile association resource allocation policies.

By introducing the weight factor ω_y, where $y \in \{1, 2, 3\}$, **MOOP1** can be reorganized as:

$$\mathbf{MOOP2} : \min_{\mathbf{x},\mathbf{n}} \max_{y=1,2,3} \left\{ \omega_y \left(U_y(\mathbf{x}, \mathbf{n}) - U_y^{\circ} \right) \right\} \tag{7.16}$$

subject to C1-C6. Without loss of generality, we set $\sum_{y=1}^{3} \omega_y = 1$. We can introduce an auxiliary variable ψ and convert MOOP2 into:

$$\mathbf{MOOP3} : \min_{\mathbf{x},\mathbf{n},\psi} \psi \tag{7.17}$$

subject to

$$C1 - C6,$$

$$C7 : \omega_1 \left\{ -\sum_{i=1}^{N_c} \sum_{j=0}^{N_r} \sum_{k=1}^{N_u} x_{i,j,k} \mathrm{PSNR}_{i,j,k} - U_1^{\circ} \right\} \leq \psi, \tag{7.18}$$

$$C8 : \omega_2 \left\{ \sum_{i=1}^{N_c} \sum_{j=0}^{N_r} \sum_{k=1}^{N_u} x_{i,j,k} \left(P_t n_{i,j,k} + P_c \right) - U_2^{\circ} \right\} \leq \psi, \tag{7.19}$$

$$C9 : \omega_3 \left\{ \sum_{i=1}^{N_c} \sum_{j=0}^{N_r} \sum_{k=1}^{N_u} x_{i,j,k} \left(\rho_m n_{i,0,k} + \rho_p n_{i,j,k} \right) - U_3^{\circ} \right\} \leq \psi. \tag{7.20}$$

The mixture of integer variables $x_{i,j,k}$ and non-integer variables $n_{i,j,k}$ make the problem very difficult to solve. We first relax the binary variable $x_{i,j,k}$ to a real value one, and introduce a new variable $\hat{n}_{i,j,k} = n_{i,j,k} x_{i,j,k}$ to denote the auxiliary network resource consumption. Then, the transformed problem is convex with respect to the variable $\hat{n}_{i,j,k}$. Assuming the existence of an interior point, then the Slaters condition is satisfied and strong duality holds. Thus, solving the dual problem is equivalent to solving the primal problem. We introduce positive dual variables $a, b, c, \lambda_i, \mu_{i,j}, \nu_k, \gamma_k, \eta_k$ and formulate the Lagrangian

of **MOOP3** as

$$\begin{aligned}
\mathcal{L}(\mathbf{\Omega}) &= \psi + a\left[\omega_1\left\{-\sum_{i=1}^{N_c}\sum_{j=0}^{N_r}\sum_{k=1}^{N_u} x_{i,j,k}\mathrm{PSNR}_{i,j,k} - U_1^{\circ}\right\} - \psi\right] \\
&+ b\left[\omega_2\left\{\sum_{i=1}^{N_c}\sum_{j=0}^{N_r}\sum_{k=1}^{N_u}\left(P_t\hat{n}_{i,j,k} + x_{i,j,k}P_c\right) - U_2^{\circ}\right\} - \psi\right] \\
&+ c\left[\omega_3\left\{\sum_{i=1}^{N_c}\sum_{j=0}^{N_r}\sum_{k=1}^{N_u}\left(\rho_m\hat{n}_{i,0,k} + \rho_p\hat{n}_{i,j,k}\right) - U_3^{\circ}\right\} - \psi\right] + \sum_{i=1}^{N_c}\lambda_i\left(\sum_{k=1}^{N_u}\hat{n}_{i,0,k} - C_i^m\right) \\
&+ \sum_{i=1}^{N_c}\sum_{j=1}^{N_r}\mu_{i,j}\left(\sum_{k=1}^{N_u}\hat{n}_{i,j,k} - C_{i,j}^p\right) + \sum_{k=1}^{N_u}\nu_k\left(\sum_{i=1}^{N_c}\sum_{j=0}^{N_r}x_{i,j,k} - 1\right) \\
&+ \sum_{k=1}^{N_u}\gamma_k\left(\mathrm{PSNR}_{min} - \mathrm{PSNR}_{i,j,k}\right) - \sum_{k=1}^{N_u}\eta_k\sum_{i=1}^{N_c}\sum_{j=0}^{N_r}\hat{n}_{i,j,k}. \qquad (7.21)
\end{aligned}$$

for $\mathbf{\Omega} = (\psi, \mathbf{x}, \hat{\mathbf{n}}, a, b, c, \boldsymbol{\lambda}, \boldsymbol{\mu}, \boldsymbol{\nu}, \boldsymbol{\gamma}, \boldsymbol{\eta})$. The corresponding dual problem is given by:

$$\max_{a,b,c,\boldsymbol{\lambda},\boldsymbol{\mu},\boldsymbol{\nu},\boldsymbol{\gamma},\boldsymbol{\eta}} \min_{\psi,\mathbf{x},\hat{\mathbf{n}}} \mathcal{L}(\psi, \mathbf{x}, \hat{\mathbf{n}}, a, b, c, \boldsymbol{\lambda}, \boldsymbol{\mu}, \boldsymbol{\nu}, \boldsymbol{\gamma}, \boldsymbol{\eta}). \qquad (7.22)$$

Let $(x_{i,j,k}^*, n_{i,j,k}^*)$ denote the optimal mobile association and resource allocation policies, so that the optimal value of ψ can be determined by:

$$\psi^* = \max_{1\le y\le 3}\left\{\omega_y\left(U_y(\mathbf{x}^*, \mathbf{n}^*) - U_y^{\circ}\right)\right\}. \qquad (7.23)$$

Known the value of ψ, by using dual decomposition technique [56], a sequence of sub-problems are solved to obtain the optimal mobile association and resource allocation for given dual variables. Then the dual variables are updated depending on the obtained mobile association and resource allocation. Iterations are completed until the convergence of the optimal dual and primal solutions is reached. Specifically, the solving process can be decomposed into following two levels:

- low-level sub-problems

$$\min_{\psi,\mathbf{x},\hat{\mathbf{n}}} \mathcal{L}(\mathbf{\Omega}) \quad \text{for } \mathbf{\Omega} = (\psi, \mathbf{x}, \hat{\mathbf{n}}, a, b, c, \boldsymbol{\lambda}, \boldsymbol{\mu}, \boldsymbol{\nu}, \boldsymbol{\gamma}, \boldsymbol{\eta})$$

- high-level master dual problem

$$\max_{a,b,c,\boldsymbol{\lambda},\boldsymbol{\mu},\boldsymbol{\nu},\boldsymbol{\gamma},\boldsymbol{\eta}} g(\mathbf{\Omega}) \quad \text{for } \mathbf{\Omega} = (\psi, \mathbf{x}, \hat{\mathbf{n}}, a, b, c, \boldsymbol{\lambda}, \boldsymbol{\mu}, \boldsymbol{\nu}, \boldsymbol{\gamma}, \boldsymbol{\eta})$$

where $g(\mathbf{\Omega}) = \min_{\psi,\mathbf{x},\hat{\mathbf{n}}} \mathcal{L}(\psi, \mathbf{x}, \hat{\mathbf{n}}, a, b, c, \boldsymbol{\lambda}, \boldsymbol{\mu}, \boldsymbol{\nu}, \boldsymbol{\gamma}, \boldsymbol{\eta})$.

7.3.1 Low-level Sub-problem

Given the dual variables, the low-level sub-problem is to solve the following minimization problem:

$$\min_{\psi,\mathbf{x},\hat{\mathbf{n}}} \mathcal{L}(\psi, \mathbf{x}, \hat{\mathbf{n}}, a, b, c, \boldsymbol{\lambda}, \boldsymbol{\mu}, \boldsymbol{\nu}, \boldsymbol{\gamma}, \boldsymbol{\eta}). \tag{7.24}$$

With KKT conditions [21], we can obtain the following optimal resource allocation for UE k:

$$\hat{n}^*_{i,0,k} = x_{i,0,k} n^*_{i,0,k} = x_{i,0,k} \left[\frac{\alpha_k (a\omega_1 + \gamma_k)}{\log 10 \times (b\omega_2 P_m + c\omega_3 \rho_m + \lambda_i - \eta_k)} \right]^+, \tag{7.25}$$

$$\hat{n}^*_{i,j,k} = x_{i,j,k} n^*_{i,j,k} = x_{i,j,k} \left[\frac{\alpha_k (a\omega_1 + \gamma_k)}{\log 10 \times (b\omega_2 P_p + c\omega_3 \rho_p + \mu_{i,j} - \eta_k)} \right]^+, \tag{7.26}$$

where $[x]^+ = \max\{0, x\}$. Substituting $n^*_{i,0,k}$ and $n^*_{i,j,k}$ back to (7.24) and taking the derivatives of the sub-problem with respect to $x_{i,0,k}$ and $x_{i,j,k}$, respectively, we can obtain

$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial x_{i,0,k}} = & -a\omega_1 \alpha_k \log_{10}\left(n^*_{i,0,k} \log_2(1 + \psi_{i,0,k})\right) + b\omega_2 \left(P_m n^*_{i,0,k} + P_c\right) + c\omega_3 \rho_m n^*_{i,0,k} \\ & + \lambda_i n^*_{i,0,k} - \gamma_k \alpha_k \log_{10}\left(n^*_{i,0,k} \log_2(1 + \psi_{i,0,k})\right) - \eta_k n^*_{i,0,k}, \end{aligned} \tag{7.27}$$

$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial x_{i,j,k}} = & -a\omega_1 \alpha_k \log_{10}\left(n^*_{i,j,k} \log_2(1 + \psi_{i,j,k})\right) + b\omega_2 \left(P_p n^*_{i,j,k} + P_c\right) + c\omega_3 \rho_p n^*_{i,j,k} \\ & + \mu_{i,j} n^*_{i,j,k} - \gamma_k \alpha_k \log_{10}\left(n^*_{i,j,k} \log_2(1 + \psi_{i,j,k})\right) - \eta_k n^*_{i,j,k}. \end{aligned} \tag{7.28}$$

In order to satisfy C6, the optimal mobile association decision for the UE k is given by

$$x^*_{i,j,k} = \begin{cases} 1; & \text{for } \{i,j\} = \arg\min \frac{\partial \mathcal{L}}{\partial x_{i,j,k}},\ \forall i,j \\ 0; & \text{otherwise.} \end{cases} \tag{7.29}$$

7.3.2 High-level Master Dual Problem

The high-level master dual problem is to obtain the dual variables by solving the dual function:

$$\max_{a,b,c,\boldsymbol{\lambda},\boldsymbol{\mu},\boldsymbol{\nu},\boldsymbol{\gamma},\boldsymbol{\eta}} g(a,b,c,\boldsymbol{\lambda},\boldsymbol{\mu},\boldsymbol{\nu},\boldsymbol{\gamma},\boldsymbol{\eta}) = \min_{\psi \mathbf{x},\hat{\mathbf{n}}} \mathcal{L}(\psi,\mathbf{x},\hat{\mathbf{n}},a,b,c,\boldsymbol{\lambda},\boldsymbol{\mu},\boldsymbol{\nu},\boldsymbol{\gamma},\boldsymbol{\eta}) \tag{7.30}$$

Since $g(\boldsymbol{\lambda},\boldsymbol{\mu},\boldsymbol{\nu},\boldsymbol{\gamma},\boldsymbol{\eta})$ is differentiable, we can obtain the dual variables by using gradient descent method. The update process is shown in (7.31)-(7.37),

$$a(t+1) = \left[a(t) - \theta_1 \left[\psi^* - \omega_1 \left\{ -\sum_{i=1}^{N_c}\sum_{j=0}^{N_r}\sum_{k=1}^{N_u} x^*_{i,j,k}\mathrm{PSNR}_{i,j,k} - U_1^{\circ} \right\} \right] \right]^+, \tag{7.31}$$

$$b(t+1) = \left[b(t) - \theta_2 \left[\psi^* - \omega_2 \left\{ \sum_{i=1}^{N_c}\sum_{j=0}^{N_r}\sum_{k=1}^{N_u} (P_t\hat{n}_{i,j,k} + x_{i,j,k}P_c) - U_2^{\circ} \right\} \right] \right]^+, \tag{7.32}$$

$$c(t+1) = \left[c(t) - \theta_3 \left[\psi^* - \omega_3 \left\{ \sum_{i=1}^{N_c}\sum_{j=0}^{N_r}\sum_{k=1}^{N_u} (\rho_m\hat{n}_{i,0,k} + \rho_p\hat{n}_{i,j,k}) - U_3^{\circ} \right\} \right] \right]^+, \tag{7.33}$$

$$\lambda_i(t+1) = \left[\lambda_i(t) - \theta_4 \left(C_i - \sum_{k=1}^{N_u} x^*_{i,0,k} n^*_{i,0,k} \right) \right]^+, \tag{7.34}$$

$$\mu_{i,j}(t+1) = \left[\mu_i(t) - \theta_5 \left(C_{i,j} - \sum_{k=1}^{N_u} x^*_{i,j,k} n^*_{i,j,k} \right) \right]^+, \tag{7.35}$$

$$\nu_k(t+1) = \left[\nu_k(t) - \theta_6 \left(1 - \sum_{i=1}^{N_c}\sum_{j=0}^{N_r} x^*_{i,j,k} \right) \right]^+, \tag{7.36}$$

$$\gamma_k(t+1) = \left[\gamma_k(t) - \theta_7 \left(\mathrm{PSNR}_{i,j,k} - \mathrm{PSNR}_{min} \right) \right]^+. \tag{7.37}$$

where t is the iteration index and $\theta_1 \sim \theta_7$ are the positive step sizes.

7.3.3 Iteration Process

For additional clarity, we classify the iteration process into an outer loop and an inner loop. Based on the optimal mobile association $\mathbf{x}^*$ and optimal resource allocation $\mathbf{n}^*$ obtained in inner loop, ψ is updated in outer loop. In inner loop, where ψ is given as a parameter, dual decomposition is employed to obtain the optimal mobile association $\mathbf{x}^*$ and optimal resource allocation $\mathbf{n}^*$ by solving a sequence of sub-problems. The inner loop stops when the dual variables converge and the outer loop stops when ψ converges.

Table 7.1: Simulation parameters

Parameter	**Settings**
mBS	57
pBS	4 per macro-cell
UE	200 per cell
Circuit Power	$P_c = 13$ dBm
Transmit Power	$P_m = 46$ dBm, $P_p = 30$ dBm
System Bandwidth	20 MHz
Noise Model and density	AWGN, -174 dBm/Hz
Base Layer PSNR	23.74 dB
Load Balancing Weight	$\rho_m = 4, \rho_p = 1$

7.4 Performance Evaluation

The simulation was set up based on 3GPP case 1 configurations specified in [22]. In a 19-cell 3-sector three-ring hexagonal network structure, one mBS is located in the center of a macro-cell and 4 pBSs are equally-distanced deployed in the overlaid pico-cells within each macro-cell, forming a two-tier heterogeneous network. UEs are uniformly distributed in the network. Standard SVC test video sequence *Foreman* in the QCIF format (176×144 pixels) with a frame rate of 15 frames/sec is used in the simulation and the PSNR for the encoded base layer is 23.74 dB. Other parameter settings are shown in Table. 7.1.

Fig. 7.1 presents the trade-off regions for the system objectives achieved by the optimal

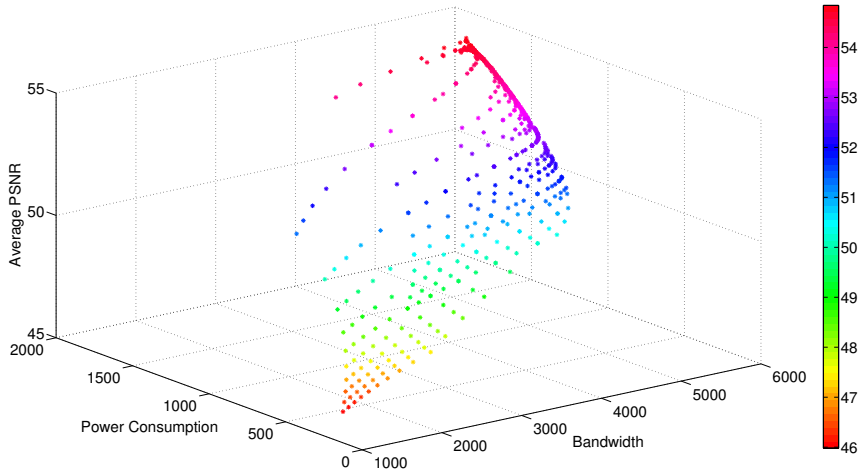

Fig. 7.1: Scatter graph

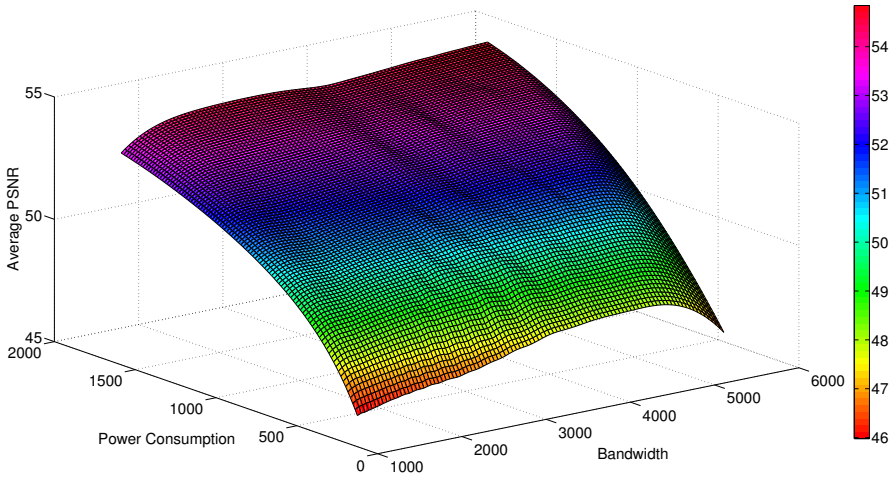

Fig. 7.2: 3-D graph

mobile association and resource allocation scheme. The asterisk markers in the scatter graph denote the Pareto optimal point. It is observed that a large portion of the trade-off region concentrates on the top of the figure, where power consumption and bandwidth consumption are relatively high. In other words, a mobile association and resource allocation scheme that has a high power consumption and bandwidth consumption also leads to a high average PSNR. Furthermore, by connecting the Pareto optimal points and applying curve fitting, we

can illustrate the trade-off region in a surface shown in Fig. 7.2, which clearly demonstrates the trade-off among three objectives. Fig. 7.3 displays the corresponding contour graph. With different values on weights ψ_1, ψ_2, ψ_3, the disparity of the average PSNR can go up to 8dB. Besides, under certain power consumption, the highest bandwidth consumption does not necessarily give the best perceived video quality PSNR. The reason lies in weight distribution for the three objectives under this scenario. Therefore, it is very significant to decide the weight distribution in the multi-objective optimization.

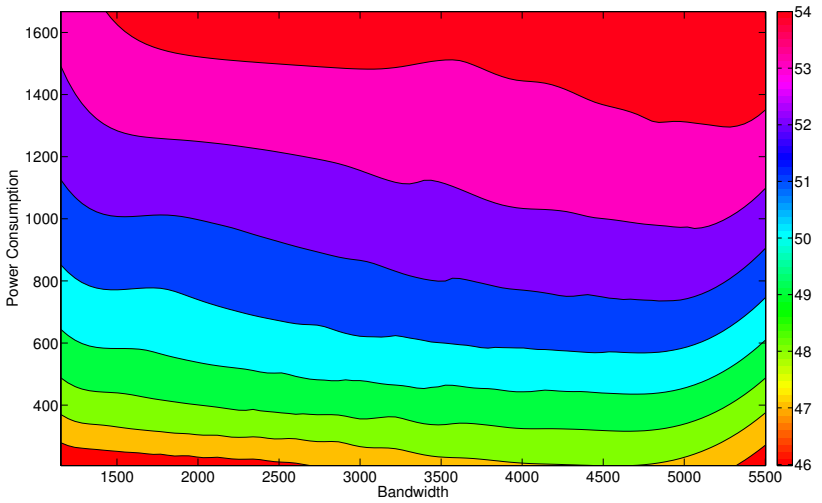

Fig. 7.3: Contour graph

Fig. 7.4(a) and Fig. 7.4(b) show the power consumption and the bandwidth consumption versus the average PSNR for different weight combinations. By setting the weight of one objective, either bandwidth consumption or power consumption, equal to zero, we can explore the trade-off of the other two objectives. It is observed that in both figures, the average PSNR goes up with the increment of the power consumption (see Fig. 7.4(a)) or the bandwidth consumption (see Fig. 7.4(b)). We can conclude that the perceived video quality improves if UEs consume more energy or more bandwidth. This is because with a higher power consumption or a higher bandwidth consumption, the channel condition is improved and the achievable data rate increases. These improvements lead to the increment of UE's PSNR, which in turn boosts the increment of the average PSNR in the entire system.

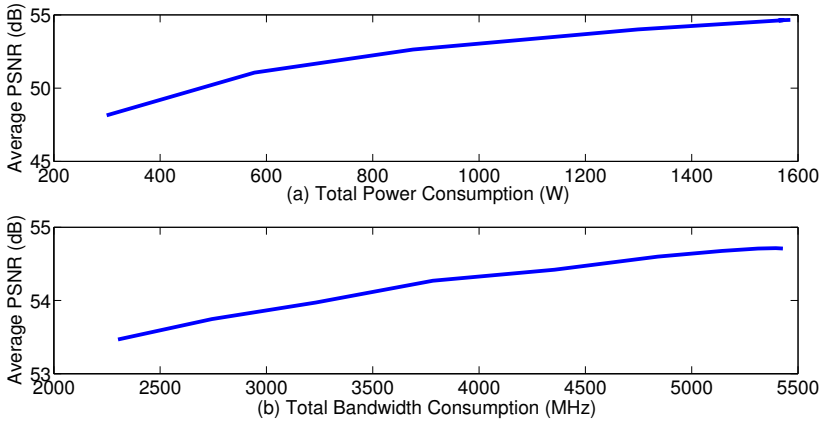

Fig. 7.4: Average PSNR at different power and bandwidth consumptions

7.5 Chapter Summary

In this chapter, we formulate a multi-objective optimization framework for the video user mobile association and resource allocation in a wireless heterogeneous network. The framework characterizes three design objectives: PSNR maximization, power consumption minimization and bandwidth consumption minimization. We solve the multi-objective optimization problem by using weighted Tchebycheff approach, and apply dual decomposition technique to obtain the optimized mobile association and resource allocation. The simulation reveals the trade-offs among three different design objectives and provides a clear understanding on how network performance compromises under conflicting design objectives.

Chapter 8

Conclusion and Future Work

8.1 Summary of Major Contributions

In this dissertation, we studied QoE-aware spectrum efficiency and energy efficiency over wireless heterogeneous networks. Specifically, we explored the benefits of cooperative transmission, precoding techniques, and NOMA technique in wireless system designs. To tackle the challenges emerging in heterogeneous networks, we applied cooperative transmission to mitigate the inter-cell and intra-cell interferences. Furthermore, we combined this with a precoding technique to enhance user performances through increasing data transmission rates. NOMA was considered to help improve spatial diversity and achieve further capacity gains. In addition, we considered the delivery of video applications over wireless heterogeneous networks. By proposing two new performance metrics, we explored the interplay of video quality-based spectrum efficiency and energy efficiency in system design.

First, we utilized inter-cell and intra-cell cooperation to mitigate the interferences between macro-cells and micro-cells. The proposed cooperative transmission scheme can increase the overall system capacity and improve cell edge user performances notably, e.g., 10 dB increment of SINRs for 30% of the total users. Further, in order to offload data traffic from macro-cells to achieve system-wise load balancing, we proposed a bias-based range expansion mobile association scheme to compensate the power disparity between macro-nodes and micro-nodes. The nonlinear precoding scheme THP was applied to cancel out the inter-user interferences and enable the cell edge user to achieve a 250% capacity gain.

Second, because of its superior spectrum efficiency, we introduced a hybrid MU-MIMO and NOMA design scheme in wireless heterogeneous networks to improve the system throughput and also to increase multi-user diversity gains by exploiting the heterogeneous nature of the supporting wireless networks. We properly chose UEs to form a NOMA pair and then

applied a precoding based MU-MIMO scheme to cancel out partial interfering signals. A brute-force search algorithm was used to solve the resource scheduling optimization problem with proportional fairness purpose. Furthermore, we proposed a cooperative NOMA framework in a multi-antenna system, where multiple users can be served concurrently. To exploit spatial diversity and mitigate the inter-user interferences, we implemented dirty paper coding on the transmitting side and successive interference cancellation on the receiving side. It was observed that given the perfect CSI, system performance was considerably improved in terms of data rates and spectrum efficiency, e.g, cooperative NOMA can achieve about 20% and 33% network throughput gain comparing to cooperation scheme and NOMA-only scheme, respectively.

Third, as multimedia services composing a huge proportion of data traffic, we considered video applications in a heterogeneous network. In order to evaluate spectrum and energy efficiency from the perspective of video quality, we proposed two new performance metrics: QSE and QEE. Then we formulated a joint mobile association and resource allocation optimization problem to explore the trade-off between QSE and QEE and their relationships to decaying factors. Furthermore, we conducted a multi-objective optimization framework to study the relationships of bandwidth consumption, power consumption, and perceived video quality. Extensive simulations were conducted to reveal the fundamental relationships among SE, EE, and QoE.

8.2 Future Work

8.2.1 Backhaul-limited Heterogeneous Networks

The majority of research on heterogeneous networks ignores the backhaul links that connect macro-nodes and micro-nodes. This might be reasonable for legacy radio access networks (RANs), given the assumption that the backhaul link is often over-provisioned (e.g., fiber). For instance, in this dissertation I considered out-of-band backhaul links and ignored their impacts on network resource consumption and mobile associations. Future networks will need to take this into consideration. The high density of small cells, and the

related network operational expenditures, suggest that backhaul links will mostly be under-provisioned and cannot be ignored [62]. Moreover, the increased backhaul signaling traffic required for CoMP [63], the backhaul resource sharing between macro-nodes and micro-nodes, as well as upcoming cloud-RAN [64] technologies, are expected to further stress the backhaul network. Thereby, as radio access technologies continue to improve, the backhaul network will emerge as a major performance bottleneck in heterogeneous networks, and mobile association schemes that ignore the backhaul load and topology will result in poor performance. Hence, in future system design and optimizations for heterogeneous networks, it will be necessary to take backhaul links into account.

8.2.2 Imperfect CSI in NOMA System

CSI on the transmitting side greatly impacts the precoding design and performance of interference alignment. Most of the existing work to design the precoder and analyze performance assumes that perfect CSI is available on both transmitting and receiving sides. However, in practice, it is frequently difficult to obtain a perfect CSI of interference channels. To solve this problem, it is important to apply channel estimation, CSI feedback, and other proper approaches. Moreover, in a real NOMA system, confronting the impact of SIC error propagation is inevitable, especially since imperfect CSI will jeopardize system performance. Hence, it is necessary and reasonable to adopt effective strategies to alleviate the impacts of SIC error propagation.

8.2.3 Hybrid User Service Strategy

In a NOMA system, performance improvement is mainly dominated by the difference of received signal strengths. Larger differences between the received signals result in higher aggregated data rates achieved by the system. The NOMA technique should only be applied if the difference in signal strengths is significant. When the difference between the received signals is marginal, user cooperation and other strategies should be considered. Therefore, a hybrid user service strategy should be studied in the future.

8.2.4 Device-to-device Communication Deployment

Recently, to facilitate green communication, device-to-device (D2D) communication was introduced to heterogeneous networks to complement cooperative transmission and short-range communication. The coexistence of cellular users and D2D pairs raised new technical challenges, e.g., interference mitigation, user grouping, resource scheduling, etc. These interesting topics drive the development of next-generation networks.

8.2.5 Dynamic Resource Scheduling in Video Communications

In this dissertation, I focused on a video application-based mobile association problem rather than a video transmission problem, i.e., the decision to accept or not accept a mobile user that will need a video connection, and how much radio resource needs to be reserved for that video connection in order to maximize system QSE and QEE. In this type of problem the decisions are made before the connection is actually set up. For this decision making, we captured the requirements for video quality in the mobile association process. Mobile association is similar to the traditional call admission control process, where there is no video traffic flowing in the network yet. In the future, it will be necessary to simulate video traffic over a heterogeneous network and study resource scheduling problems during video transmission, where the system dynamically assigns network resources and determines video content delivery according to end-user requirements.

References

[1] R. Q. Hu and Y. Qian, *Heterogeneous Cellular Networks*. John Wiley & Sons, Ltd., 2013.

[2] ——, *Resource Management for Heterogeneous Networks in LTE Systems*. Springer, 2014.

[3] Q. Li, R. Hu, Y. Qian, and G. Wu, "Cooperative communications for wireless networks: techniques and applications in lte-advanced systems," *Wireless Communications, IEEE*, vol. 19, no. 2, pp. –, April 2012.

[4] A. Ghosh, N. Mangalvedhe, R. Ratasuk, B. Mondal, M. Cudak, E. Visotsky, T. Thomas, J. G. Andrews, P. Xia, H. S. Jo *et al.*, "Heterogeneous cellular networks: From theory to practice," *Communications Magazine, IEEE*, vol. 50, no. 6, pp. 54–64, 2012.

[5] Whitepaper c11-481360, "Cisco visual networking index: Forecast and methodology," Jun. 2011.

[6] Ericsson, "A discussion on some technology components for lte-advanced," 3GPP TSG-RAN WG1 #53bis R1-082024, May 2008.

[7] Nokia Siemens Networks, "Further dl comp phase1 simulation results," 3GPP TSG-RAN WG1 #65bis R1-111276, May 2011.

[8] Huawei, HiSilicon, "Final dl comp jp performance evaluation of phase 1," 3GPP TSG-RAN WG1 #65bis R1-111338, May 2011.

[9] D. Choi, D. Lee, and J. H. Lee, "Resource allocation for comp with multiuser mimo-ofdma," *Vehicular Technology, IEEE Transactions on*, vol. 60, no. 9, pp. 4626–4632, Nov 2011.

[10] X. Huang, Y. Mao, F. Wu, and S. Leng, "Resource allocation for qos-aware ofdma cellular networks with cooperative relaying," *WTOC*, vol. 10, no. 1, pp. 12–23, Jan. 2011. [Online]. Available: http://dl.acm.org/citation.cfm?id=2001165.2001167

[11] 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 J. Wirel. Commun. Netw.*, vol. 2009, pp. 24:1–24:14, Jan. 2009. [Online]. Available: http://dx.doi.org/10.1155/2009/475281

[12] C. Mueller, A. Klein, and B. Raaf, "A coordinated resource allocation algorithm for infrastructure-based relay networks," *EURASIP J. Adv. Sig. Proc.*, vol. 2009, 2009. [Online]. Available: http://dx.doi.org/10.1155/2009/630964

[13] A. Damnjanovic, J. Montojo, Y. Wei, T. Ji, T. Luo, M. Vajapeyam, T. Yoo, O. Song, and D. Malladi, "A survey on 3gpp heterogeneous networks," *Wireless Communications, IEEE*, vol. 18, no. 3, pp. 10–21, June 2011.

[14] Huawei, "Discussion on relay in comp," 3GPP TSG-RAN WG1 meeting #65bis, Jun. 2009.

[15] Qualcomm Europe, "Range expansion for efficient support of heterogeneous networks," 3GPP TSG-RAN WG1 #54bis R1-083813, Sep. 2008.

[16] K. Son, S. Chong, and G. Veciana, "Dynamic association for load balancing and interference avoidance in multi-cell networks," *Wireless Communications, IEEE Transactions on*, vol. 8, no. 7, pp. 3566–3576, July 2009.

[17] L. Tassiulas and S. Sarkar, "Maxmin fair scheduling in wireless networks," in *INFOCOM 2002. Twenty-First Annual Joint Conference of the IEEE Computer and Communications Societies. Proceedings. IEEE*, vol. 2, 2002, pp. 763–772 vol.2.

[18] P. Viswanath, D. Tse, and R. Laroia, "Opportunistic beamforming using dumb antennas," *Information Theory, IEEE Transactions on*, vol. 48, no. 6, pp. 1277–1294, Jun 2002.

[19] J. Sun, E. Modiano, and L. Zheng, "Wireless channel allocation using an auction algorithm," *Selected Areas in Communications, IEEE Journal on*, vol. 24, no. 5, pp. 1085–1096, May 2006.

[20] Z.-Q. Luo and W. Yu, "An introduction to convex optimization for communications and signal processing," *Selected Areas in Communications, IEEE Journal on*, vol. 24, no. 8, pp. 1426–1438, Aug 2006.

[21] S. Boyd and V. Vandenberghe, *Convex Optimization*. Cambridge University Press, 2004.

[22] 3GPP TR36.814, "Further advancements for e-utra physical layer aspects," v9.0.0, Mar. 2010.

[23] Y. Zhou, J. Wang, and M. Sawahashi, "Downlink transmission of broadband ofcdm systems-part i: hybrid detection," *Communications, IEEE Transactions on*, vol. 53, no. 4, pp. 718–729, April 2005.

[24] Huawei, "Practical analysis of comp coordinated beamforming," 3GPP TSG-RAN WG1 #58bis R1-093036, Aug. 2009.

[25] Q. Li, R. Hu, Y. Qian, and G. Wu, "Intracell cooperation and resource allocation in a heterogeneous network with relays," *Vehicular Technology, IEEE Transactions on*, vol. 62, no. 4, pp. 1770–1784, May 2013.

[26] Y. Yu, Q. Hu, C. Bontu, and Z. Cai, "Mobile association and load balancing in a cooperative relay cellular network," *Communications Magazine, IEEE*, vol. 49, no. 5, pp. 83–89, May 2011.

[27] B. Wang, B. Li, and M. Liu, "A novel precoding method for joint processing in comp," in *Network Computing and Information Security (NCIS), 2011 International Conference on*, vol. 1, May 2011, pp. 126–129.

[28] H. Kim and Y. Han, "A proportional fair scheduling for multicarrier transmission systems," *Communications Letters, IEEE*, vol. 9, no. 3, pp. 210–212, March 2005.

[29] A. L. Stolyar, "On the asymptotic optimality of the gradient scheduling algorithm for multiuser throughput allocation," *Oper. Res.*, vol. 53, no. 1, pp. 12–25, Jan. 2005. [Online]. Available: http://dx.doi.org/10.1287/opre.1040.0156

[30] Y. Saito, Y. Kishiyama, A. Benjebbour, T. Nakamura, A. Li, and K. Higuchi, "Non-orthogonal multiple access (noma) for cellular future radio access," in *Vehicular Technology Conference (VTC Spring), 2013 IEEE 77th*, June 2013, pp. 1–5.

[31] Z. Ding, P. Fan, and V. Poor, "Impact of user pairing on 5g non-orthogonal multiple access downlink transmissions," *Vehicular Technology, IEEE Transactions on*, vol. PP, no. 99, pp. 1–1, 2015.

[32] Y. Saito, A. Benjebbour, Y. Kishiyama, and T. Nakamura, "System-level performance evaluation of downlink non-orthogonal multiple access (noma)," in *Personal Indoor and Mobile Radio Communications (PIMRC), 2013 IEEE 24th International Symposium on*, Sept 2013, pp. 611–615.

[33] M. Costa, "Writing on dirty paper (corresp.)," *Information Theory, IEEE Transactions on*, vol. 29, no. 3, pp. 439–441, May 1983.

[34] Y. Zheng and C. Xiao, "Simulation models with correct statistical properties for rayleigh fading channels," *Communications, IEEE Transactions on*, vol. 51, no. 6, pp. 920–928, June 2003.

[35] E. Falkenauer, *Genetic Algorithms and Grouping Problems*. John Wiley & Sons, Ltd., 1998.

[36] J. Fan, G. Li, Q. Yin, B. Peng, and X. Zhu, "Joint user pairing and resource allocation for lte uplink transmission," *Wireless Communications, IEEE Transactions on*, vol. 11, no. 8, pp. 2838–2847, August 2012.

[37] Whitepaper c11-520862, "Cisco visual networking index: Global mobile data traffic forecast update, 2012-2017," Feb. 2013.

[38] M. Ismail, W. Zhuang, and S. Elhedhli, "Energy and content aware multi-homing video transmission in heterogeneous networks," *Wireless Communications, IEEE Transactions on*, vol. 12, no. 7, pp. 3600–3610, July 2013.

[39] M. Ismail and W. Zhuang, "Mobile terminal energy management for sustainable multi-homing video transmission," *Wireless Communications, IEEE Transactions on*, vol. 13, no. 8, pp. 4616–4627, Aug 2014.

[40] W. Song and W. Zhuang, "Performance analysis of probabilistic multipath transmission of video streaming traffic over multi-radio wireless devices," *Wireless Communications, IEEE Transactions on*, vol. 11, no. 4, pp. 1554–1564, April 2012.

[41] W. Dinkelbach, "On nonlinear fractional programming," *Management Science*, vol. 13, no. 7, pp. pp. 492–498, 1967. [Online]. Available: http://www.jstor.org/stable/2627691

[42] Y. Xu, R. Q. Hu, L. Wei, and G. Wu, "Qoe-aware mobile association and resource allocation over wireless heterogeneous networks," in *IEEE Global Communications Conference, GLOBECOM 2014, Austin, TX, USA, December 8-12, 2014*, 2014, pp. 4695–4701. [Online]. Available: http://dx.doi.org/10.1109/GLOCOM.2014.7037549

[43] M. Chen, M. Ponec, S. Sengupta, J. Li, and P. A. Chou, "Utility maximization in peer-to-peer systems with applications to video conferencing," *IEEE/ACM Trans. Netw.*, vol. 20, no. 6, pp. 1681–1694, Dec. 2012. [Online]. Available: http://dx.doi.org/10.1109/TNET.2012.2201166

[44] Y. F. Z. K. Hu and H. He, *Internet Resource Pricing Models*. Springer, 2013.

[45] G. Li, Z. Xu, C. Xiong, C. Yang, S. Zhang, Y. Chen, and S. Xu, "Energy-efficient wireless communications: tutorial, survey, and open issues," *Wireless Communications, IEEE*, vol. 18, no. 6, pp. 28–35, December 2011.

[46] C. Shannon, "A mathematical theory of communication," *Bell System Technical Journal, The*, vol. 27, no. 3, pp. 379–423, July 1948.

[47] G. Lim and L. J. C. Jr., "Energy-efficient cooperative relaying in heterogeneous radio access networks," *IEEE Wireless Commun. Letters*, vol. 1, no. 5, pp. 476–479, 2012. [Online]. Available: http://dx.doi.org/10.1109/WCL.2012.070312.120366

[48] R. F. B. F. W. J. Oliver, D. W. Lozier and W. C. Charles, *NIST Handbook of Mathematical Functions*. Cambridge University Press, 2005.

[49] S. Cui, A. Goldsmith, and A. Bahai, "Energy-constrained modulation optimization," *Wireless Communications, IEEE Transactions on*, vol. 4, no. 5, pp. 2349–2360, Sept 2005.

[50] Q. Li, Y. Xu, R. Hu, and G. Wu, "Pricing-based distributed mobile association for heterogeneous networks with cooperative relays," in *Communications (ICC), 2012 IEEE International Conference on*, June 2012, pp. 5326–5331.

[51] Y. Xu and R. Hu, "Optimal intra-cell cooperation in the heterogeneous relay networks," in *Global Communications Conference (GLOBECOM), 2012 IEEE*, Dec 2012, pp. 4120–4125.

[52] Y. Xu, R. Hu, Q. Li, and Y. Qian, "Optimal intra-cell cooperation with precoding in wireless heterogeneous networks," in *Wireless Communications and Networking Conference (WCNC), 2013 IEEE*, April 2013, pp. 761–766.

[53] Q. C. Li, R. Q. Hu, Y. Xu, and Y. Qian, "Optimal fractional frequency reuse and power control in the heterogeneous wireless networks," *IEEE Transactions on Wireless Communications*, vol. 12, no. 6, pp. 2658–2668, 2013. [Online]. Available: http://dx.doi.org/10.1109/TWC.2013.050313.120160

[54] R. Marler and J. Arora, "Survey of multi-objective optimization methods for engineering," *Structural and Multidisciplinary Optimization*, vol. 26, no. 6, pp. 369–395, 2004. [Online]. Available: http://dx.doi.org/10.1007/s00158-003-0368-6

[55] D. Palomar and M. Chiang, "A tutorial on decomposition methods for network utility maximization," *Selected Areas in Communications, IEEE Journal on*, vol. 24, no. 8, pp. 1439–1451, Aug. 2006.

[56] D. Ng, E. Lo, and R. Schober, "Wireless information and power transfer: Energy efficiency optimization in ofdma systems," *Wireless Communications, IEEE Transactions on*, vol. 12, no. 12, pp. 6352–6370, December 2013.

[57] YUV Video Sequences. [Online]. Available: http://trace.eas.asu.edu/yuv/index.html

[58] J. Klaue, B. Rathke, and A. Wolisz, *EvalVid A Framework for Video Transmission and Quality Evaluation.* Springer Berlin Heidelberg, 2003. [Online]. Available: http://dx.doi.org/10.1007/978-3-540-45232-4_16

[59] Whitepaper c11-520862, "Cisco visual networking index: Global mobile data traffic forecast update, 2014-2019," Feb. 2015.

[60] R. T. Marler and J. S. Arora, "Survey of multi-objective optimization methods for engineering," *Structural and Multidisciplinary Optimization*, vol. 26, pp. 369–395, Apr. 2004.

[61] H. Schwarz, D. Marpe, and T. Wiegand, "Overview of the scalable video coding extension of the h.264/avc standard," *Circuits and Systems for Video Technology, IEEE Transactions on*, vol. 17, no. 9, pp. 1103–1120, Sep. 2007.

[62] *Backhaul technologies for small cells.* Small Cell Forum, 2014.

[63] J. Lee, Y. Kim, H. Lee, B. L. Ng, D. Mazzarese, J. Liu, W. Xiao, and Y. Zhou, "Coordinated multipoint transmission and reception in lte-advanced systems," *Communications Magazine, IEEE*, vol. 50, no. 11, pp. 44–50, November 2012.

[64] P. Rost, C. Bernardos, A. Domenico, M. Girolamo, M. Lalam, A. Maeder, D. Sabella, and D. Wubben, "Cloud technologies for flexible 5g radio access networks," *Communications Magazine, IEEE*, vol. 52, no. 5, pp. 68–76, May 2014.

Appendices

Appendix A

Proof of Theorem 1

***Theorem 1*:** q^* can be obtained if and only if

$$\begin{aligned} \mathbf{P4}: F(q^*) &= \min_{\mathbf{n}\in S_{\mathbf{n}}, \mathbf{x}\in S_{\mathbf{x}}} Q(\mathbf{n},\mathbf{x}) - q^* D(\mathbf{n},\mathbf{x}) \\ &= Q(\mathbf{n}^*,\mathbf{x}^*) - q^* D(\mathbf{n}^*,\mathbf{x}^*) \\ &= 0 \end{aligned}$$

for any $Q(\mathbf{n},\mathbf{x}) > 0$ and $D(\mathbf{n},\mathbf{x}) > 0$.

Proof. We prove the sufficiency and necessity of the theorem separately.

a) Sufficiency: Suppose we have optimal solutions $\mathbf{n}^*$ and $\mathbf{x}^*$. From (6.31), we have

$$\begin{aligned} Q(\mathbf{n},\mathbf{x}) - q^* D(\mathbf{n},\mathbf{x}) &\geq \min_{\mathbf{n}\in S_{\mathbf{n}}, \mathbf{x}\in S_{\mathbf{x}}} Q(\mathbf{n},\mathbf{x}) - q^* D(\mathbf{n},\mathbf{x}) \\ &= Q(\mathbf{n}^*,\mathbf{x}^*) - q^* D(\mathbf{n}^*,\mathbf{x}^*) \\ &= 0. \end{aligned} \tag{A.1}$$

Because of the positivity characteristic of $D(\mathbf{n},\mathbf{x})$, we can obtain

$$q^* \leq \frac{Q(\mathbf{n},\mathbf{x})}{D(\mathbf{n},\mathbf{x})} \quad \text{and} \quad q^* = \frac{Q(\mathbf{n}^*,\mathbf{x}^*)}{D(\mathbf{n}^*,\mathbf{x}^*)}. \tag{A.2}$$

Here, q^* is the optimal solution (minimum) of **P3** with the optimal solutions $\mathbf{n}^*$ and $\mathbf{x}^*$. Sufficiency proof completes.

b) Necessity: Assume q^* is the optimal solution of **P3** and it is always a positive value due to the definition. Then we have

$$q^* = \frac{Q(\mathbf{n}^*,\mathbf{x}^*)}{D(\mathbf{n}^*,\mathbf{x}^*)} = \min_{\mathbf{n}\in S_{\mathbf{n}}, \mathbf{x}\in S_{\mathbf{x}}} \frac{Q(\mathbf{n},\mathbf{x})}{D(\mathbf{n},\mathbf{x})} \leq \frac{Q(\mathbf{n},\mathbf{x})}{D(\mathbf{n},\mathbf{x})}. \tag{A.3}$$

Hence, $Q(\mathbf{n}, \mathbf{x}) - q^* D(\mathbf{n}, \mathbf{x}) \geq 0$ for all $\mathbf{n} \in S_{\mathbf{n}}, \mathbf{x} \in S_{\mathbf{x}}$. And its value is equal to zero when $\mathbf{n}$ and $\mathbf{x}$ approach the optimal solutions. We can write

$$F(q^*) = Q(\mathbf{n}^*, \mathbf{x}^*) - q^* D(\mathbf{n}^*, \mathbf{x}^*) = 0. \tag{A.4}$$

Necessity proof completes. □

Appendix B

Convergence Proof of Algorithm 3

Before we proceed with the proof, we first introduce the following two lemmas. Then with the help of these two lemmas, we prove that q decreases in each iteration step and converges to its optimum with sufficient iterations, and $F(q)$ converges to zero so that the optimality is satisfied.

***Lemma 1*:** $F(q) = \min_{\mathbf{n}\in S_{\mathbf{n}},\mathbf{x}\in S_{\mathbf{x}}} Q(\mathbf{n},\mathbf{x}) - qD(\mathbf{n},\mathbf{x})$ is strictly monotonically decreasing in q, e.g., $F(q_{k+1}) > F(q_k)$, if $q_k > q_{k+1}$.

Proof. Given $q_k > q_{k+1}$, suppose $(\mathbf{n}_k^*, \mathbf{x}_k^*)$ and $(\mathbf{n}_{k+1}^*, \mathbf{x}_{k+1}^*)$ are the optimal solutions of $F(q_k)$ and $F(q_{k+1})$, respectively. It is known that $D(\mathbf{n},\mathbf{x}) > 0$, Then

$$\begin{aligned} F(q_{k+1}) &= Q(\mathbf{n}_{k+1}^*, \mathbf{x}_{k+1}^*) - q_{k+1}D(\mathbf{n}_{k+1}^*, \mathbf{x}_{k+1}^*) \\ &> Q(\mathbf{n}_{k+1}^*, \mathbf{x}_{k+1}^*) - q_kD(\mathbf{n}_{k+1}^*, \mathbf{x}_{k+1}^*) \\ &\geq \min_{\mathbf{n}_k\in S_{\mathbf{n}},\mathbf{x}_k\in S_{\mathbf{x}}} Q(\mathbf{n}_k,\mathbf{x}_k) - q_kD(\mathbf{n}_k,\mathbf{x}_k) \\ &= Q(\mathbf{n}_k^*, \mathbf{x}_k^*) - q_kD(\mathbf{n}_k^*, \mathbf{x}_k^*) \\ &= F(q_k). \end{aligned} \tag{B.1}$$

□

***Lemma 2*:** For arbitrary $\mathbf{n}_k \in S_{\mathbf{n}}, \mathbf{x}_k \in S_{\mathbf{x}}$, and $q_k = \frac{Q(\mathbf{n}_k,\mathbf{x}_k)}{D(\mathbf{n}_k,\mathbf{x}_k)}$, we have $F(q_k) \leq 0$.

Proof. Since $q_k = \frac{Q(\mathbf{n}_k,\mathbf{x}_k)}{D(\mathbf{n}_k,\mathbf{x}_k)}$, we have $Q(\mathbf{n}_k,\mathbf{x}_k) = q_kD(\mathbf{n}_k,\mathbf{x}_k)$. Then,

$$\begin{aligned} F(q_k) &= \min_{\mathbf{n}\in S_{\mathbf{n}},\mathbf{x}\in S_{\mathbf{x}}} Q(\mathbf{n},\mathbf{x}) - q_kD(\mathbf{n},\mathbf{x}) \\ &\leq Q(\mathbf{n}_k,\mathbf{x}_k) - q_kD(\mathbf{n}_k,\mathbf{x}_k) = 0 \end{aligned} \tag{B.2}$$

□

In order to prove convergence, we denote $(\mathbf{n}_k, \mathbf{x}_k)$ as the optimal solution of $F(q)$ at the kth iteration, and q_k as the corresponding value. Then at the $(k+1)$th iteration, q_{k+1} is updated by

$$q_{k+1} = \frac{Q(\mathbf{n}_k, \mathbf{x}_k)}{D(\mathbf{n}_k, \mathbf{x}_k)} \tag{B.3}$$

in Algorithm 3. Note that neither q_k or q_{k+1} equal to q^* so that the ***Lemma 2*** holds.

It is easy to know that $F(q_k) < 0$ and $F(q_{k+1}) < 0$ and $Q(\mathbf{n}_k, \mathbf{x}_k) = q_{k+1} D(\mathbf{n}_k, \mathbf{x}_k)$. Then we have the following relationship

$$\begin{aligned} F(q_k) &= Q(\mathbf{n}_k, \mathbf{x}_k) - q_k D(\mathbf{n}_k, \mathbf{x}_k) \\ &= q_{k+1} D(\mathbf{n}_k, \mathbf{x}_k) - q_k D(\mathbf{n}_k, \mathbf{x}_k) \\ &= (q_{k+1} - q_k) D(\mathbf{n}_k, \mathbf{x}_k) < 0. \end{aligned} \tag{B.4}$$

for $D(\mathbf{n}_k, \mathbf{x}_k) > 0$. Thus, we have $q_{k+1} < q_k$, which means that q decreases in each iteration. When the number of iterations $k \to \infty$, we have $\lim_{k\to\infty} q_k = q^*$, and because $F(q_k)$ is monotonically decreasing in p, we have $\lim_{k\to\infty} F(q_k) = F(q^*) = 0$. Based on ***Theorem 1***, the optimality is satisfied. If q_k does not converge to q^*, then there should exist another $q^\diamond$ which is $\lim_{k\to\infty} q_k = q^\diamond > q^*$ and make $\lim_{k\to\infty} F(q_k) = F(q^\diamond) = 0$. This is contradicted to ***Lemma 1*** that $F(q^*) > F(q^\diamond)$, if $q^\diamond > q^*$. Therefore, convergence to the optimal value q is guaranteed.

Appendix C

Proof of Quasi-convexity of P6 with Respect to $\hat{n}$

For single variable function, the proof of quasi-convexity can be based on the following proposition [21].

Proposition: A single variable function $f(x)$ is quasi-convex if and only if either

- it is nondecreasing, or
- it is nonincreasing, or
- there exists x^* such that $f(x)$ is nonincreasing for $x < x^*$ and nondecreasing for $x > x^*$.

Proof. For a given q, **P6** can be written as a objective function of variable $\hat{\mathbf{n}}$, which is expressed as

$$\mathbf{P6}: \min_{\hat{\mathbf{n}}} U(\hat{\mathbf{n}}) = Q(\hat{\mathbf{n}}) - qD(\hat{\mathbf{n}}), \tag{C.1}$$

where $Q(\hat{\mathbf{n}})$ and $D(\hat{\mathbf{n}})$ are both continuous functions on variable $\hat{\mathbf{n}}$. Then, we take a partial derivative of the objective function $U(\hat{\mathbf{n}})$ with respect to $\hat{n}$, which yields in (C.2):

$$\begin{aligned} U(\hat{\mathbf{n}})' &= \frac{\partial U(\hat{\mathbf{n}})}{\partial \hat{n}} = \frac{\partial Q(\hat{\mathbf{n}})}{\partial \hat{n}} - q\frac{\partial D(\hat{\mathbf{n}})}{\partial \hat{n}} \\ &= \omega_1\beta\left(\frac{\xi}{N_{ref}} + \frac{P_t}{\zeta}\right)P_{t-1}^{\beta-1}(\hat{n}) + \omega_2\theta\rho W_{t-1}^{\theta-1}(\hat{n}) - q\frac{\alpha}{\ln 10 \times \hat{n}} \\ &= \frac{\ln 10 \times \hat{n} \times \left[\omega_1\beta\left(\frac{\xi}{N_{ref}} + \frac{P_t}{\zeta}\right)P_{t-1}^{\beta-1}(\hat{n}) + \omega_2\theta\rho W_{t-1}^{\theta-1}(\hat{n})\right] - q \times \alpha}{\ln 10 \times \hat{n}}, \end{aligned} \tag{C.2}$$

Here, $\hat{n} > 0$ so that $A(\hat{n}) > 0$ and $B(\hat{n}) > 0$. It is easy to observe that when $\hat{n} \to 0^+$, $A(\hat{n}) - q \times \alpha < 0$, thereby we have $U(\hat{n})'|_{0^+} < 0$. Similarly, when $\hat{n} \to \infty$, we have $A(\hat{n}) - q \times \alpha > 0$ and $U(\hat{n})'|_{\infty} > 0$, correspondingly. Due to the continuity on $\hat{n}$, the objective function **P6** is first monotonically nonincreasing and then monotonically nondecreasing with

respect to $\hat{n}$. Thus, according to the aforementioned **Proposition**, we can conclude that **P6** is quasi-convex with respect to $\hat{n}$. □

Vita

Yiran Xu

Yiran Xu was born in China on January 7th, 1987. He obtained the B.S. degree in Electronics and Information Engineering from Huazhong University of Science and Technology, Wuhan, China, in 2009, and M.S. degree in Electrical Engineering From New York University Tandon School of Engineering, Brooklyn, New York, in 2011. Since July 2011, he has been a Ph.D. candidate at Electrical and Computer Engineering Department in Utah State University, Logan, Utah, under the supervision of Prof. Rose Qingyang Hu. His research interests include cross-layer design in energy efficient and spectrum efficient design in heterogeneous networks and quality-aware video transmission. In summer 2015, he interned at EMC Corporation. He is a two-time recipient of the competitive Student Travel Award at IEEE Globecom Conference in 2012 and 2014, respectively.

List of publications

- Y. Xu, R. Q. Hu, Y. Qian, and T. Znati, "Video Quality-based Spectrum and Energy Efficient Mobile Association in Wireless Heterogeneous Networks," *IEEE Trans. Commun.*, to appear in 2016.
- Q. Li, R. Q. Hu, Y. Xu, and Y. Qian, "Optimal fractional frequency reuse and power control in the heterogeneous wireless networks," *IEEE Trans. Wireless Commun.*, vol. 12, no.6, pp. 2658-2668, Jun. 2013.
- Y. Xu, H, Sun, and R. Q. Hu, "Hybrid MU-MIMO and Non-orthogonal Multiple Access Design in Wireless Heterogeneous Networks," submitted to *IEEE ICC 2016*.

- H. Sun, Y. Xu, and R. Q. Hu, "D2D Communications Underlying Non-orthogonal Multiple Access in a Downlink MU-MIMO Cellular Networks," submitted to *IEEE VTC Spring 2016*.

- Y. Xu, H. Sun, R. Q. Hu, and Y. Qian, "Cooperative Non-orthogonal Multiple Access in Heterogeneous Networks," in *Proc. IEEE Globecom 2015*, San Diego, Dec. 2015.

- Y. Xu, R. Q. Hu, Y. Qian, and T. Znati, "Tradeoffs in Video Transmission over Wireless Heterogeneous Networks: Energy, Bandwidth and QoE," in *Proc. IEEE ICC 2015*, London, Jun. 2015.

- Y. Xu, R. Q. Hu, L. Wei, and G. Wu, "QoE-aware Mobile Association and Resource Allocation Over Wireless Heterogeneous Networks," in *Proc. IEEE Globecom 2014*, pp. 4695-4701, Austin, Dec. 2014.

- L. Wei, Y. Xu, R. Q. Hu, and Y. Qian, "An Algebraic Framework for Mobile Association in Wireless Heterogeneous Networks," in *Proc. IEEE Globecom 2013*, Atlanta, Dec. 2013.

- Y. Xu, R. Q. Hu, Q. Li, and Y. Qian, "Optimal Intra-Cell Cooperation With Precoding in Wireless Heterogeneous Networks," in *Proc. IEEE WCNC 2013*, pp. 761-766, Shanghai, Apr. 2013.

- Y. Xu, and R. Q. Hu, "Optimal Intra-cell Cooperation in the Heterogeneous Relay Network," in *Proc. IEEE Globecom 2012*, pp. 4120-4125, Los Angeles, Dec. 2012.

- Q. Li, Y. Xu, R. Q. Hu, and G. Wu, "Pricing-based mobile association for cooperative wireless heterogeneous networks," in *Proc. IEEE ICC 2012*, pp. 5326-5331, Ottawa, Jun. 2012.

- E. Lu, Y. Xu, and I-Tai Lu, "Efficient MMSE design for joint MIMO processing in analog network coding schemes," in *Proc. IEEE ICNC 2012*, pp. 267-271, Maui, Jan. 2012.