Resource Allocation with Multi-cell Coordination
in Wireless Networks

by

Binglai Niu

B.Sc., Fudan University, Shanghai, China, 2008
M.Sc., University of Alberta, Alberta, Canada, 2010

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

in

The Faculty of Graduate and Postdoctoral Studies

(Electrical and Computer Engineering)

THE UNIVERSITY OF BRITISH COLUMBIA
(Vancouver)

September 2015

© Binglai Niu, 2015
Abstract

To meet the growing demand of mobile data service with limited radio resources, the cellular architecture has evolved from single-cell networks towards multi-cell networks. In multi-cell networks, the spectrum is reused by multiple adjacent cells to increase the spectral efficiency. As a trade-off, interference is introduced among the cells, which limits the achievable data rates for users who experience significant inter-cell interference. In this thesis, multi-cell coordination is applied to mitigate interference, and several resource allocation mechanisms are proposed to improve the system performance for various multi-cell networks. First, a downlink scheduling mechanism is proposed for a multi-cell multiple-input multiple-output (MIMO) network. This mechanism dynamically selects the users to be scheduled and the corresponding MIMO transmission strategy to optimize a utility function. Both centralized and distributed algorithms are developed, and an efficient rate adjustment method is proposed to improve the system throughput when the channel state information (CSI) is imperfect. Next, a network configuration mechanism is developed for two-tier macro-femto networks. In this mechanism, coordination is applied for different network configuration processes such that access control, spectrum allocation and power management are performed sequentially at the base stations and users, respectively. This mechanism is modeled as a multi-stage decision making process and the desired decisions are obtained using a multi-level optimization approach. Finally, coordination among multiple service providers for resource sharing is studied in cloud-based radio access networks (C-RANs). A multi-timescale resource sharing
mechanism is designed. This mechanism employs a threshold-based policy to control the inter-cell interference, and defines a new metric for providing resource sharing guarantee for each service provider. It consists of resource allocation processes that are performed at different time scales to deal with traffic demand variation. The proposed mechanism addresses the issue of user mobility by employing a mobility prediction method when optimizing the resource sharing decisions. The performance of the mechanisms proposed in this thesis are evaluated via computer simulations. It is shown that these mechanisms substantially improve the performance for the corresponding multi-cell networks.
Preface

This thesis is the original work by the author, Binglai Niu. Chapters 2-4 encompass work that has been published, accepted for publication or is under review. The corresponding papers are co-authored by Dr. Vincent W. S. Wong, who supervised the author of this thesis through the research. The papers corresponding to Chapter 2 are also co-authored by Dr. Robert Schober, who provided valuable comments on these works. The following publications describe the work completed in this thesis.

**Journal Papers Published and Accepted**


**Journal Paper Submitted**


**Conference Papers Published**

IEEE Global Communications Conference (GLOBECOM), Houston, Texas, December 2011.


# Table of Contents

Abstract ......................................................... ii

Preface ............................................................ iv

Table of Contents ............................................... vi

List of Tables .................................................... x

List of Figures ................................................... xi

List of Acronyms ................................................. xiv

Acknowledgments ................................................ xvi

1 Introduction .................................................... 1
   1.1 Background and Motivation ................................. 2
      1.1.1 Multi-cell Wireless Networks ........................ 2
      1.1.2 Resource Allocation in Multi-cell Networks ........ 6
      1.1.3 Coordination in Multi-cell Networks ................. 8
      1.1.4 Motivation ........................................... 10
   1.2 Summary of Results and Contributions .................... 14
   1.3 Thesis Organization ..................................... 16
# Table of Contents

## 2 Scheduling with Coordination for Multi-cell MIMO Networks

2.1 Introduction .............................................. 18
2.2 Related Works ........................................... 19
2.3 System Model and Problem Formulation ....................... 20
   2.3.1 System Model ...................................... 20
   2.3.2 System Stability ..................................... 22
   2.3.3 MIMO Transmission Strategies ....................... 23
   2.3.4 Problem Formulation .................................. 26
2.4 Dynamic Scheduling Mechanisms ............................. 28
   2.4.1 Centralized Dynamic Scheduling Mechanism ............. 28
   2.4.2 Distributed Dynamic Scheduling Mechanism ............. 32
   2.4.3 Scheduling with Imperfect CSI ........................ 36
2.5 Performance Evaluation .................................... 38
2.6 Summary .................................................. 46

## 3 Network Configuration for Two-tier Heterogeneous Networks

3.1 Introduction .............................................. 47
3.2 Related Works ........................................... 48
3.3 System Model ............................................. 49
   3.3.1 Utility Function ..................................... 53
   3.3.2 Network Configuration Mechanism ....................... 56
   3.3.3 Multi-level Optimization Problem ....................... 58
3.4 Analysis of the Multi-level Optimization Problem .......... 61
   3.4.1 Power Management at Users ........................... 61
   3.4.2 Resource Allocation at Base Stations ................... 63
   3.4.3 Channel Allocation at the Scheduler .................... 66
## Table of Contents

3.4.4 Access Control at FBSs ........................................... 72
3.4.5 Procedures of the Proposed Mechanism ............................ 74
3.4.6 Complexity of the Proposed Algorithms ............................ 75
3.5 Performance Evaluation .............................................. 76
3.6 Summary ............................................................. 84

4 Resource Sharing for Cloud-based Radio Access Networks ............ 85
4.1 Introduction .......................................................... 85
4.2 Related Works ........................................................ 86
4.3 System Model ......................................................... 87
  4.3.1 User-Centric Resource Sharing Scheme ........................... 89
  4.3.2 Resource Allocation Problem at the Operator ....................... 93
4.4 Efficient Resource Allocation Algorithm ................................ 95
  4.4.1 Problem Transformation ........................................... 95
  4.4.2 Increment-based Greedy Allocation Algorithm ...................... 97
4.5 Multi-Timescale Resource Sharing Mechanism ........................ 101
4.6 Further Extension .................................................... 104
  4.6.1 Dynamic Resource Sharing for Uplink ............................ 105
  4.6.2 Revenue Maximization for On-demand Service ..................... 108
4.7 Performance Evaluation ............................................... 110
  4.7.1 Efficiency of the Proposed Algorithm ............................ 111
  4.7.2 Effect of Weighting Factors ...................................... 112
  4.7.3 Effect of Interference Threshold $\xi$ ................................ 112
  4.7.4 Effect of Dynamic Resource Guarantee ........................... 113
  4.7.5 Comparison with Proportional Spectrum Sharing Mechanism ... 115
  4.7.6 Effect of Traffic Variation ...................................... 116
## List of Tables

<table>
<thead>
<tr>
<th></th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.1</td>
<td>List of key notations used in this chapter.</td>
<td>51</td>
</tr>
<tr>
<td>3.2</td>
<td>Average running time versus number of channels.</td>
<td>80</td>
</tr>
<tr>
<td>4.1</td>
<td>Average running time versus number of users per service provider.</td>
<td>111</td>
</tr>
</tbody>
</table>
List of Figures

2.1 Multi-cell MIMO systems. (a) Two-cell case; (b) Three-cell case. Each cell $i \in \mathcal{I}$ has $K_i$ users. .......................................................... 20
2.2 Two-cell model: the average utility versus transmission power $P_T$ with $d = 1$ km and $N_e = 3$. ................................................. 39
2.3 Two-cell model: the average utility versus the distance $d$ for $P_T = 20$ dBm and $N_e = 3$. ................................................. 41
2.4 Two-cell model: the average utility versus the number of cell-edge users $N_e$ for $P_T = 20$ dBm and $d = 1$ km. ................................. 42
2.5 Two-cell model: the average utility versus transmission power $P_T$ for $d = 1$ km and $N_e = 3, 5$. ................................. 43
2.6 Two-cell model: the average utility versus transmission power $P_T$ for $\xi_0 = 0.05$. ................................................. 44
2.7 Two-cell model: the average utility versus $\xi_0$ for $P_T = 20, 30$ dBm. ................................. 44
2.8 Three-cell model: the average utility versus transmission power $P_T$ for perfect CSI. ................................. 45
2.9 Three-cell model: the average utility versus transmission power $P_T$ for imperfect CSI for $\xi_0 = 0.05$. ................................. 46

3.1 Two-tier macro-femto system model: the MBS has a set of subscribed MUs and each FBS has a set of subscribed FUs. ................................. 50
3.2 User’s satisfactory function $f_i(\overline{R}_{ij})$ with respect to the average data rate $\overline{R}_{ij}$. $R_{ij}^{\text{min}}$ is the minimum rate requirement, $C$ is a constant, and $\Phi_i(\cdot)$ is a concave function defined in $[R_{ij}^{\text{min}}, +\infty)$. .............................. 54

3.3 Five-stage network configuration mechanism: Stages I and II correspond to access control process, Stages III and IV are the resource allocation process. Stage V is the power management process. .............................. 57

3.4 Total utility of the system versus total number of channels $N$, with $r = 40$ m and $d = 150$ m. ................................................................. 78

3.5 System throughput versus total number of channels $N$, with $r = 40$ m and $d = 150$ m. ................................................................. 79

3.6 Total utility of the system versus radius of the region, with $N = 30$, and $d = 150$ m. ................................................................. 80

3.7 Total utility of the system versus $d$, with $N = 30$ and $r = 40$ m. ............ 81

3.8 Total utility versus number of MUs moved into the femtocell region with $N = 30$, $d = 150$ m and $r = 40$ m. .............................. 82

3.9 Energy consumption and user satisfaction versus $\beta$, with $N = 28$, $d = 120$ m and $r = 40$ m. .............................. 83

4.1 A cloud-based radio access network. Multiple service providers lease the infrastructure and radio resource from a network operator. The network operator creates vRAN for each service provider by assigning a number of RRHs and the corresponding channel resources. Each RRH can serve users from different service providers. .............................. 88
4.2 (a) Multi-timescale resource sharing framework: Global resource allocation performed every $T_G$ time slots, and local resource allocation performed every $T_L$ time slots. (b) Mobility prediction: predicting the locations of mobile users every $\Delta t$ time slots during the next $T_L$ time slots. 102

4.3 System throughput versus number of users per service provider. 110

4.4 System throughput with different weighting factors versus number of users. 112

4.5 System throughput versus different interference threshold $\xi$. 113

4.6 System throughput versus number of users per service provider. 114

4.7 System throughput versus number of users per service provider. 115

4.8 System throughput versus time slot. 116

4.9 System throughput versus average speed of mobile users. 118

4.10 System throughput versus prediction error. 119
## List of Acronyms

<table>
<thead>
<tr>
<th>Acronym</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>3GPP</td>
<td>3rd Generation Partnership Project</td>
</tr>
<tr>
<td>AWGN</td>
<td>Additive White Gaussian Noise</td>
</tr>
<tr>
<td>BBU</td>
<td>Baseband Unit</td>
</tr>
<tr>
<td>CA</td>
<td>Carrier Aggregation</td>
</tr>
<tr>
<td>CAPEX</td>
<td>Capital Expenses</td>
</tr>
<tr>
<td>CoMP</td>
<td>Coordinated Multi-Point</td>
</tr>
<tr>
<td>C-RANs</td>
<td>Cloud-based Radio Access Networks</td>
</tr>
<tr>
<td>CSI</td>
<td>Channel State Information</td>
</tr>
<tr>
<td>CSMA</td>
<td>Carrier Sense Multiple Access</td>
</tr>
<tr>
<td>FBSs</td>
<td>Femto Base Stations</td>
</tr>
<tr>
<td>FUs</td>
<td>Femtocell Users</td>
</tr>
<tr>
<td>GPS</td>
<td>Global Positioning System</td>
</tr>
<tr>
<td>IBGA</td>
<td>Increment-Based Greedy Allocation</td>
</tr>
<tr>
<td>i.i.d.</td>
<td>independent and identically distributed</td>
</tr>
<tr>
<td>LMMSE</td>
<td>Linear Minimum Mean Square Error</td>
</tr>
<tr>
<td>LTE-Advanced</td>
<td>Long Term Evolution Advanced</td>
</tr>
<tr>
<td>MBS</td>
<td>Macro Base Station</td>
</tr>
<tr>
<td>MIMO</td>
<td>Multiple-Input Multiple-Output</td>
</tr>
<tr>
<td>MILP</td>
<td>Mixed-Integer Linear Programming</td>
</tr>
</tbody>
</table>
### List of Acronyms

<table>
<thead>
<tr>
<th>Acronym</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>MUs</td>
<td>Macrocell Users</td>
</tr>
<tr>
<td>OFDM</td>
<td>Orthogonal Frequency-Division Multiplexing</td>
</tr>
<tr>
<td>OPEX</td>
<td>Operational Expenses</td>
</tr>
<tr>
<td>QoS</td>
<td>Quality of Service</td>
</tr>
<tr>
<td>RAU</td>
<td>Radio Access Unit</td>
</tr>
<tr>
<td>RANs</td>
<td>Radio Access Networks</td>
</tr>
<tr>
<td>RRHs</td>
<td>Remote Radio Heads</td>
</tr>
<tr>
<td>SIC</td>
<td>Successive Interference Cancellation</td>
</tr>
<tr>
<td>SVD</td>
<td>Singular Value Decomposition</td>
</tr>
<tr>
<td>vRANs</td>
<td>virtual Radio Access Networks</td>
</tr>
<tr>
<td>WSRM</td>
<td>Weighted Sum-Rate Maximization</td>
</tr>
</tbody>
</table>
Acknowledgments

First of all, I would like to express my deepest sense of gratitude to my supervisor, Dr. Vincent Wong, for providing me valuable advice and guidance in every stage of my research. Throughout the training under his guidance, I established solid analytical and problem solving skills, which prepared me for any challenge in the future career.

I shall extend my sincere thanks to Dr. Robert Schober for his enormous advice and comments during the collaboration of my first project. Thanks to his help, I got continuous progress in my research and built a solid foundation for the following projects.

I am deeply grateful for my colleagues in my lab: Dr. Vahid Shah-Mansouri, Dr. Derrick Wing Kwan Ng, Dr. Keivan Ronasi, Dr. Pedram Samadi, Dr. Man Hon Cheung, Dr. Hamed Shah-Mansouri, Dr. Homa Eghbali, Shaobo Mao, Suyang Duan, Enxin Yao, Bojiang Ma, Zehua Wang, Peiran Wu, Jun Zhu, Hao Ma, Boya Song, and Yanan Sun. I thank them for offering me a comfortable working environment and providing valuable comments on my research work.

Most importantly, I feel indebted to my parents, for their everlasting love and support not only in years of doing PhD but also throughout my entire life. I owe my deepest gratitude to my wife. I would not reach so far without her understanding, support and love.

This work was supported by the Natural Sciences and Engineering Research Council (NSERC) of Canada and Four Year Doctoral Fellowship program from the University of British Columbia.
Chapter 1

Introduction

With the popularity of mobile devices and the explosive growth in mobile Internet services, wireless communication systems are expected to provide high speed and high quality data services for a variety of wireless applications [1]. Due to the scarcity of radio resources (e.g., frequency spectrum, transmission power), efficient transmission strategies and resource allocation mechanisms are needed to achieve high data rates with limited radio resources. Traditional transmission strategies and resource allocation mechanisms are designed based on a single-cell deployment, where each cell is treated independently. Such design requires careful frequency planning so that adjacent cells use different frequency bands to avoid inter-cell interference. However, the single-cell deployment has low spectral efficiency and cannot fulfill the growing demand for high speed wireless data services. As a result, multi-cell processing has been proposed as a promising solution to improve spectral efficiency and system throughput [2]. With multi-cell processing, adjacent base stations (or access points) reuse the same frequency spectrum to serve their intended users. The radio resources are utilized more efficiently via dynamic resource allocation among multiple cells. High data rates are achievable by exploiting concurrent transmissions through advanced techniques such as orthogonal frequency-division multiplexing (OFDM) and multiple-input multiple-output (MIMO) spatial multiplexing [3]. Despite the advantages, multi-cell processing also brings challenges in designing efficient resource allocation mechanisms. Since the same frequency spectrum is reused by several adjacent cells, users may experience inter-cell interference, especially for those who are located at the edge of adjacent cells. The existence
of inter-cell interference may degrade the signal quality in each cell and further limit the system performance [4]. Therefore, when designing resource allocation mechanisms, coordination among multiple cells is required to mitigate interference and to optimize the system performance.

In this thesis, resource allocation mechanisms with multi-cell coordination are proposed for different types of multi-cell wireless networks, including multi-cell MIMO networks, two-tier macro-femto networks, and cloud-based radio access networks (C-RANs). These mechanisms aim to reduce inter-cell interference and improve system performance by using various optimization techniques, such as stochastic network optimization, multi-level optimization, and convex optimization. The rest of this chapter is organized as follows. Section 1.1 provides an overview of the background and motivation of this thesis. The main contributions and results are summarized in Section 1.2. The organization of this thesis is described in Section 1.3.

1.1 Background and Motivation

1.1.1 Multi-cell Wireless Networks

Multi-cell is a spectral efficient architecture in wireless communication systems. A multi-cell network consists of a cluster of adjacent base stations, which utilize the same licensed spectrum to serve their intended users. Compared to the conventional single-cell based communication systems, multi-cell networks have the following advantages:

- Efficient spectrum utilization: In conventional cellular systems, the licensed spectrum is divided into several bands which are reused following a specific pattern, i.e., adjacent cells are allocated orthogonal channels and each cell can only use part of the spectrum. In multi-cell networks, each cell can utilize the entire spectrum.
Although interference exists due to the universal frequency reuse, by applying interference mitigation techniques [5], multi-cell networks achieve higher system throughput than conventional cellular systems. Therefore, the spectrum resources are utilized more efficiently in multi-cell networks.

- Dynamic resource allocation: In conventional cellular systems, resource allocation is performed in each cell independently with fixed partition of spectrum resources. In multi-cell networks, with advanced signal processing techniques, such as OFDM and carrier aggregation (CA), the entire spectrum can be divided into a number of subcarriers and dynamically allocated to users in each cell to optimize the system performance. The flexibility of resource allocation makes it easy for joint optimization over multiple cells to improve the system performance.

- Advanced transmission strategies: In multi-cell networks, advanced transmission strategies, such as coordinated multipoint (CoMP) transmission [6] and interference alignment [7], can be used through coordination among multiple cells. These advanced transmission strategies are shown to be able to mitigate inter-cell interference and to achieve high data rates.

Depending on the hierarchy of cell deployment, multi-cell networks can be classified into single-tier networks and multi-tier heterogeneous networks. A single-tier network consists of cells of the same type, i.e., all cells have the same size, and their corresponding base stations have the same transmission capabilities, such as the number of antennas and transmission power. Examples of single-tier multi-cell networks include the cellular networks with macrocell deployment and C-RANs. In multi-tier heterogeneous networks, small cells, such as picocells or femtocells, are deployed within the conventional macrocell architecture to expand the coverage area. An example of this type of networks is the two-tier macro-femto networks, where femtocells are deployed within the macrocell to enhance
Chapter 1. Introduction

the data rates for indoor users. Such two-tier networks can be deployed in urban residential areas to meet the high volume of traffic demand. This thesis studies both types of multi-cell networks, including single-tier multi-cell MIMO networks, two-tier macro-femto networks, and C-RANs, which are introduced as follows.

- Multi-cell MIMO networks: The development of multi-cell MIMO networks has been incorporated in designing the fourth generation wireless communication systems, such as the Long Term Evolution Advanced (LTE-Advanced) system [8]. In multi-cell MIMO networks, both base stations and user devices are equipped with multiple antennas. High data rates are achievable for point-to-point communication by exploiting concurrent transmissions of multiple data streams using MIMO processing. However, since the same frequency spectrum is reused by adjacent cells, inter-cell interference is experienced at the user devices (in the downlink) or at the base stations (in the uplink), which affects the performance of MIMO transmission strategies [9]. Therefore, interference management becomes a major challenge when deploying multi-cell MIMO networks. With multiple antennas, advanced MIMO processing techniques, such as coordinated beamforming [10], joint transmission [11], and interference alignment can be applied to mitigate interference and to improve the system performance.

- Two-tier macro-femto networks: In two-tier macro-femto networks, femto base stations (FBSs) are placed within the existing macrocell infrastructure to extend the coverage area of the macro base station (MBS) and increase the system capacity [12]. In particular, in urban residential area, installing FBSs in houses and apartments has been considered as an efficient approach to improve the indoor signal quality [13]. FBSs are low-cost low-power base stations named home nodeB or home eNodeB in the 3rd Generation Partnership Project (3GPP) specification.
Chapter 1. Introduction

FBSs are installed by users and can serve a small number of subscribers at the operator’s licensed spectrum [14]. By deploying FBSs, home users’ devices can experience better indoor signal quality and longer battery life due to the close proximity. The network operator can also offload the data traffic at the MBSs to FBSs in order to improve the overall system performance [15]. There are several challenges in deploying two-tier macro-femto networks. First, since both MBSs and FBSs use the same licensed spectrum, a radio resource management mechanism is needed to manage the cross-tier interference between macrocells and femtocells, as well as the inter-cell interference among the femtocells [16]. Second, it is shown that the overall system performance is improved if FBSs serve some macrocell users (MUs) who are close to them. However, an incentive mechanism is required to motivate the FBSs to perform such operation.

• C-RANs: Deploying C-RANs has been proposed recently as a novel cost-efficient solution to address the increasing demand of wireless data traffic [17, 18]. In C-RAN architecture, the baseband unit (BBU) and radio access unit (RAU) are decoupled. The BBUs, which are responsible for processing baseband signals, are placed in a cloud-based data center for centralized control and processing [19]. There are three major components in a C-RAN: a data center consisting of a pool of BBUs, a large number of remote radio heads (RRHs) each deployed in a small cell, and a high-speed optical backhaul that connects the data center and the RRHs. The data center is responsible for performing advanced centralized processing such as resource allocation and interference management. The RRHs equipped with simple radio transceivers are responsible for delivering (or receiving) data packets to (or from) the mobile users. There are several advantages in deploying a C-RAN, such as reducing capital expenses (CAPEX) and operational expenses (OPEX) for system upgrade.
and maintenance [20], and improving spectral efficiency via centralized interference control and CoMP transmission [21, 22]. An application of C-RAN is to support multiple service providers which lease radio resources from a single network operator to serve their subscribed users.

In this thesis, multi-cell MIMO networks are studied in Chapter 2, and two-tier macro-femto networks are analyzed in Chapter 3. Finally, C-RANs are studied in Chapter 4.

1.1.2 Resource Allocation in Multi-cell Networks

Resource allocation plays an important role in improving the system performance of multi-cell networks. The desired resource allocation mechanism aims to allocate the available resources, such as spectrum, antenna elements, or transmission power, to the base stations and users in order to optimize the system utility or throughput. Several aspects are considered when designing an efficient resource allocation mechanism, which are introduced as follows.

- Access control: The access control (or user association) process in multi-cell networks is to determine which user should be served by which base station at a given time. For example, in two-tier macro-femto networks where FBSs operate at hybrid access mode, an MU can be served either by the MBS or by an FBS. However, an FBS may have limited resources or transmission capability. Therefore, an access control scheme is desirable to determine the set of users to be served for each FBS to optimize the overall system performance.

- Channel allocation: In multi-cell networks, the available spectrum is divided into a number of frequency channels (or subcarriers) with fixed bandwidth. In the channel allocation process, the central controller or network operator allocates the available
channels to each base station to optimize the system performance. Several factors should be considered in designing the channel allocation scheme, such as reducing interference, balancing traffic load, and maximizing throughput. The desired channel allocation decisions can be obtained by solving network-wide optimization problems.

- **Scheduling**: In a time-slotted system, the time frame is divided into slots. In each time slot, the scheduling process selects a set of users to be served with the objective to optimize certain performance metrics such as stability, throughput, and fairness [23–25]. There are several scheduling mechanisms proposed in the literature, such as opportunistic scheduling, throughput-optimal scheduling, and round robin scheduling. Since the scheduling mechanism is usually designed based on the achievable data rates, in multi-cell networks, the selection of interference management technique should be considered when designing efficient scheduling mechanisms. For example, in multi-cell MIMO networks, different MIMO transmission strategies may achieve different data rates when a user experiences different levels of interference. The desired scheduling mechanism should determine which user to be served as well as which transmission strategy to be used.

- **Power management**: The power management process determines the transmission power at base stations (in the downlink) or users (in the uplink). The optimal power can be determined in a centralized manner by solving optimization problems with the objective of maximizing system utility or throughput. It can also be determined at each base station (or user) independently, with individual objectives, such as maximizing the achievable data rate of a link. Since transmission power has a direct impact on the level of inter-cell interference, the power management mechanism may affect the overall system performance significantly.
Chapter 1. Introduction

Traditional resource allocation mechanisms are designed based on a single-cell processing technique, where the radio resources are allocated to each cell independently without coordination or cooperation among the base stations [26]. However, such design requires network-wide frequency planning and orthogonal channel allocation among the adjacent base stations [27], which results in inefficient utilization of the spectrum. To improve the spectral efficiency, multi-cell processing techniques have been proposed, which incorporate joint processing among the base stations. Therefore, when designing resource allocation mechanisms for multi-cell networks, coordination among the base stations are required in order to apply multi-cell processing techniques to mitigate interference and to optimize the system performance.

1.1.3 Coordination in Multi-cell Networks

Multi-cell networks are interference limited due to the universal reuse of the frequency spectrum, which leverage the use of coordination among base stations to mitigate interference. The coordination approaches can generally be summarized into two directions: coordination for transmission and coordination for resource allocation.

- Coordination for transmission: In this approach, adjacent base stations exchange information and coordinate with each other in designing the transmission strategy to mitigate interference. Some of the transmission strategies require sharing only the channel state information (CSI), such as coordinated beamforming and interference alignment. In these strategies, each base station serves its own user, while the precoding and decoding matrices are coordinated among the base stations so that inter-cell interference is reduced or aligned at the same spatial dimension at the receiver. These strategies do not impose restrictions over the backhaul capacity or buffer size at each base station. Other transmission strategies require sharing of
the users’ data, such as joint transmission [11] and network MIMO [28]. These approaches require multiple base stations serve one user by jointly transmitting the same data. These strategies can achieve higher data rates at the cost of a larger buffer size at each base station and having stringent capacity requirements of the backhaul links to exchange those data.

• Coordination for resource allocation: In this approach, base station coordinate with each other when allocating the radio resources such as spectrum and transmission power. Examples of this approach include fractional frequency reuse [29–31] and power control [32, 33]. With fractional frequency reuse, the available spectrum is divided into two bands. The cellular cell is also partitioned into two regions: the cell-center region and the cell-edge region. For the cell-center regions, one frequency band is deployed with a reuse factor of 1, where the entire frequency band can be reused in any cell-center region. For the cell-edge regions, a frequency reuse factor of 3 is used where this band is divided into 3 orthogonal channels and different channels are allocated to adjacent cell-edge regions. This approach reduces inter-cell interference due to the long distance between cell-center regions. However, careful frequency planning is required for the cell-edge regions. With power control, base stations (or users) coordinate with each other to control their transmission power in order to reduce interference generated to each other. Such coordination process is typically modeled as a game, and game theoretical approaches can be applied to find a steady state solution [34]. Power control is typically used in heterogeneous networks. For example, when the same channel is used by both the macrocell and femtocells, the FBSs should restrict their transmission power so that the transmission in femtocell does not affect the on-going transmission in the macrocell.
These two coordination approaches can be applied simultaneously to improve the performance of multi-cell networks. For example, when designing rate-based scheduling mechanisms, different transmission strategies can be selected based on the coordination among base stations to optimize the data rates.

1.1.4 Motivation

As discussed in Sections 1.1.2 and 1.1.3, coordination plays an important role in designing efficient resource allocation mechanisms for multi-cell networks. The objective of the research in this thesis is to improve the performance of multi-cell networks by designing resource allocation mechanisms with various coordination approaches. The research work is presented in accordance with the evolution of the system model, from single-tier networks to multi-tier networks, and from network with a single service provider to network with multiple service providers.

This thesis begins with applying coordination for transmission in single-tier multi-cell networks in Chapter 2, where a scheduling mechanism with transmission strategy selection is proposed to optimize the utility for multi-cell MIMO networks. The motivation of this chapter is two-fold. First, MIMO transmission strategies achieve different performance under different levels of interference. In multi-cell MIMO networks, high data rates are achievable at the receivers by enabling concurrent transmissions of multiple data streams using spatial multiplexing. However, the performance of spatial multiplexing degrades in the presence of interference. Specifically, spatial multiplexing does not achieve high data rates for cell-edge users in multi-cell networks. Recently, interference alignment has emerged as an effective technique to suppress inter-cell interference [35–40]. Interference alignment allows multiple transmitters to jointly design precoding and decoding matrices so that multiple concurrent interference-free
transmissions are achievable by sacrificing one spatial dimension to accommodate interference at each transmitter-receiver pair. Since an additional spatial dimension is sacrificed for accommodating interference rather than transmitting user data, interference alignment may be outperformed by spatial multiplexing when the inter-cell interference is insignificant. Therefore, in a multi-cell network with users who experience various levels of inter-cell interference, the performance can be optimized by allowing the base stations to choose between different MIMO transmission strategies. Second, the diversity of transmission strategy brings new challenge to designing scheduling mechanisms in multi-cell MIMO networks. On one hand, scheduling mechanisms are typically designed to optimize system utility or throughput, which depends on the transmission strategy selected for each link. On the other hand, implementing the MIMO transmission strategy, such as interference alignment, requires coordination and the knowledge of scheduling decisions, i.e., which users are scheduled for transmission simultaneously and their corresponding CSI. Therefore, it is desirable to jointly consider the user scheduling process and the transmission strategy selection process to optimize the system performance, which motivates the research in Chapter 2.

Next, a network configuration mechanism is designed for a two-tier macro-femto network in Chapter 3. The motivation of the research in this chapter is explained as follows. As discussed in Section 1.1.1, deploying FBSs within the macrocell networks is a promising solution to provide high speed data services in urban residential areas. Since FBSs are user deployed, typically they are configured to operate in closed access mode and only serve the authorized subscribers, referred as femtocell users (FUs). However, it has been shown that allowing non-subscribers, i.e., MUs, to access FBSs can improve the overall system performance when the MUs experience poor signal quality from the MBS [41]. As a result, hybrid access mode has been proposed to deal with this issue. In hybrid
access mode, FBSs may grant access to MUs with limited resources while serving their subscribed FUs with higher priority [42]. This access mode aims to seek a balance between improving the overall system performance and preserving the quality of service (QoS) at the FUs, and has become a popular choice in designing novel network configuration mechanisms [43]. However, since an FBS is owned by its FUs, incentives should be provided to motivate the operation of hybrid access mode. This issue has been studied in the literature and several pricing-based mechanisms have been proposed [44–48]. These pricing-based mechanisms are developed based on specific business models, which may not be available in the existing systems. In practice, an MU may pay access fee to the service provider instead of making separate payments to FBSs. Therefore, it is interesting to study how to encourage the FBSs to adopt hybrid access without changing the business model, which motivates the research in Chapter 3. In addition to providing incentives for using hybrid access mode at the FBSs, an efficient access control scheme is needed to determine the set of MUs to be served at each FBS. In practice, the access control, resource allocation, and power management processes are performed at different individuals (i.e., the network operator, base stations and users) which may have different optimization objectives. Such decision making processes are usually correlated and an efficient network configuration mechanism is needed to coordinate these processes to improve the overall system performance. Therefore, optimization of access control, resource allocation and power management should be jointly considered when designing the network configuration mechanism. These issues are addressed in Chapter 3.

The system models considered in Chapters 2 and 3 correspond to the scenario where only a single service provider exists in the entire network. It is also interesting to study the problem where multiple service providers share the radio resources as well as the network
infrastructure. Resource sharing has several potential use cases [49, 50]. For example, C-RAN is a candidate to provide network sharing for multiple service providers. Service providers which do not own the network can lease wireless resources from a C-RAN operator to serve their subscribed users. This can help the network operator generate more revenue, and can also reduce CAPEX and OPEX for deploying network infrastructure at each service provider. To enable network sharing, a C-RAN operator needs to dynamically allocate the radio resources among service providers, which can be achieved via virtualization [51].

The basic idea of virtualization in radio access networks (RANs) is to create virtual radio access networks (vRANs) based on the physical infrastructure, where each vRAN shares a certain amount of wireless resources such that the operation in one vRAN does not affect other vRANs. In the literature, several principles have been proposed for designing virtualization mechanisms [52], such as to provide service isolation among the vRANs, to offer flexible customization capability, and to utilize radio resources efficiently. The 3GPP specification also introduces some features for resource sharing mechanisms in their study items, which include supporting on-demand service requests and adapting to varying traffic load conditions [53].

Designing resource sharing mechanism in C-RANs is challenging due to several reasons. First, to improve spectral efficiency among densely deployed small cells in a C-RAN, centralized interference control techniques (such as CoMP) are used, which usually require information exchange and coordination among the RRHs to guarantee the performance [54]. Second, in C-RANs, users’ mobility may trigger handoff across small cells frequently, which makes it challenging for providing interference control and service isolation among service providers. Moreover, due to the variation of users’ traffic demand and locations, the amount of resources required at a small cell by each service provider changes, which requires dynamic update of vRANs accordingly. These issues have not
been fully addressed, which motivates the research in Chapter 4.

### 1.2 Summary of Results and Contributions

This thesis studies several resource allocation problems in different types of multi-cell wireless networks, which includes multi-cell MIMO networks, two-tier macro-femto networks, and C-RANs. The research work is divided into three chapters. The contributions in each chapter are as follows:

- Chapter 2 studies downlink scheduling problem with the consideration of transmission strategy selection in multi-cell MIMO networks. In the considered system, each base station maintains a finite data buffer and users experience different levels of inter-cell interference. Two MIMO transmission strategies, spatial multiplexing and interference alignment, are used to optimize users’ data rates. A discrete-time stochastic optimization problem is formulated which aims to jointly select a user and the corresponding transmission strategy for each base station in order to maximize the overall system utility while stabilizing all data queues. Based on a stochastic optimization approach, a centralized dynamic scheduling algorithm with transmission strategy selection is developed. To reduce the communication overhead, a distributed scheduling algorithm is proposed which only requires limited message exchange among the base stations. When the CSI is imperfect at the base stations, the data rate calculated assuming perfect CSI is not accurate and outages may occur during the transmission. To address this issue, a rate adjustment scheme is introduced to improve the transmission success probability. It is shown that this approach improves the performance significantly compared to the case without rate adjustment. Simulation results show that the proposed scheduling mechanism is superior compared to scheduling mechanisms with a single
transmission strategy. The proposed distributed scheduling algorithm achieves a performance that is close to that of the centralized scheduling algorithm, and both algorithms achieve better performance than weighted sum-rate maximization scheduling. The work in Chapter 2 is published in [55, 56].

- Chapter 3 analyzes the uplink resource allocation problem for a two-tier macro-femto network, where an MBS and a cluster of adjacent FBSs together serve mobile users with different QoS requirements. In this system, base stations and users are individual decision makers with different optimization objectives. A five-stage network configuration mechanism is proposed which consists of access control, channel allocation, and power management processes. In this mechanism, the central scheduler, base stations and users make decisions sequentially to optimize their own objectives. This decision making process is modeled as a multi-level optimization problem. This optimization problem is analyzed in a bottom-up manner, and efficient algorithms are developed to find the solution for each level. Simulation results show that the proposed mechanism achieves higher system utility and throughput compared to the mechanisms with closed access scheme or a topology-based access control with an orthogonal channel allocation scheme. The results also show that the proposed mechanism can provide incentives for using hybrid access mode at the FBSs. The work in Chapter 3 is published in [57, 58].

- Chapter 4 addresses a resource sharing problem in C-RANs, where several service providers lease the radio resources from a network operator to serve their subscribers. A user-centric resource sharing mechanism is proposed, where the network operator jointly determines the resource to be allocated to each service provider as well as the user admission and association decisions. To guarantee service isolation, a threshold-
Chapter 1. Introduction

A novel metric is introduced to determine the minimum aggregate data rate to be provided for each service provider based on the corresponding subscribed users’ QoS requirements and their maximum achievable data rates. The proposed mechanism also employs a mobility prediction approach to estimate the locations of users in a short period and uses this information for traffic demand estimation and interference control. To assist the resource sharing process, an efficient resource allocation algorithm is proposed, which is much faster than the standard technique. A multi-timescale resource allocation framework is proposed to address the issue of traffic variation and user mobility. This framework consists of a global resource allocation process and several local resource allocation processes to deal with the variation of the network status periodically, which is more efficient than performing network-wide optimization only. Through extensive simulations, it is shown that the proposed mechanism achieves efficient resource utilization and service isolation among the service providers under various network situations with different traffic loads. The work in Chapter 4 has been submitted to a journal [59].

1.3 Thesis Organization

The rest of the thesis is organized as follows. In Chapter 2, we propose a downlink scheduling mechanism with transmission strategy selection for a multi-cell MIMO network. We formulate a discrete-time stochastic optimization problem and design both centralized and distributed algorithms to achieve a suboptimal solution. We also discuss an imperfect CSI scenario and propose a rate adjustment scheme to improve the system performance when the CSI is imperfect. In Chapter 3, we study uplink resource allocation for a two-tier heterogeneous network. We propose a multi-stage network
configuration mechanism and formulate it as a multi-level optimization problem. We apply a bottom-up approach and develop efficient algorithms to solve the optimization problem in each level. We also provide complexity analysis of the proposed algorithms and present the corresponding simulation results. In Chapter 4, we address resource sharing problem in C-RANs. We propose a multi-timescale resource sharing mechanism where the network operator jointly determines the resource to be allocated to each service provider as well as the user admission and association decisions in two different timescales. We provide extensive simulation studies of the proposed mechanism. We also discuss possible extensions for uplink communication and revenue maximization under different pricing schemes.

Each of the main chapters in this thesis is self-contained and included in separate journal articles or conference papers. The notations are defined separately for each chapter.
Chapter 2

Scheduling with Coordination for Multi-cell MIMO Networks

2.1 Introduction

Multi-cell MIMO networks have the potential to achieve high data rates by exploring advanced MIMO transmission strategies. However, MIMO transmission strategies, such as spatial multiplexing and interference alignment, have different performance under various levels of interference. To fully explore the benefits of multiple antennas, coordination among the base stations are required to select the transmission strategy to optimize the system performance. In this chapter, we design a dynamic scheduling mechanism with transmission strategy selection for multi-cell MIMO networks. There are several issues to be considered when designing scheduling mechanisms for this system. First, the scheduling mechanism should guarantee the stability of this system, which means that the data queue at each base station should not grow to infinity. Second, a proper transmission strategy should be selected for each user to manage interference and to optimize system performance. Moreover, the scheduling mechanism should provide a certain level of fairness among the users. We first design and analyze the scheduling mechanism with perfect CSI at the base stations, and then describe its extension to the realistic scenario of imperfect CSI. We formulate a discrete-time stochastic optimization problem which aims to jointly select a user and the corresponding transmission strategy for each base station in order to maximize the
overall system utility while stabilizing all data queues. Based on a stochastic optimization approach, we develop both centralized and distributed dynamic scheduling mechanisms with transmission strategy selection.

The rest of this chapter is organized as follows. In Section 2.2, we provide a summary of related works. In Section 2.3, we describe the system model and formulate the joint scheduling and transmission strategy selection problem. In Section 2.4, we develop the centralized and distributed dynamic scheduling mechanisms, and discuss the impact of imperfect CSI. Simulation results are presented in Section 2.5, and summary is given in Section 2.6.

### 2.2 Related Works

Scheduling plays an important role in improving the system performance [60, 61]. For time slotted systems, where the time frame is divided into slots, in each time slot, the downlink scheduling mechanism selects a set of users to be served with the objective to optimize certain performance metrics such as stability, throughput, and fairness [23–25]. In the literature, several scheduling mechanisms have been proposed for MIMO systems under different queueing models, such as the infinitely backlogged model. The scheduling objective is to maximize the throughput or a utility function [33, 62]. However, most of the existing scheduling mechanisms are designed for a single-cell scenario where inter-cell interference is not considered. The mechanism in [63] considers scheduling in the presence of inter-cell interference, but only studies a non-cooperative scenario where base stations schedule their transmissions independently without any interference mitigation technique. Joint scheduling, power allocation, and precoder design has been studied for multi-cell networks in [64]. However, the mechanism proposed in [64] considers only spatial multiplexing transmission for the infinitely backlogged model.
Chapter 2. Scheduling with Coordination for Multi-cell MIMO Networks

2.3 System Model and Problem Formulation

In this section, we first introduce the system model, and then we formulate a dynamic scheduling problem considering transmission strategy selection, system stability, and fairness among users.

2.3.1 System Model

We consider the downlink of a MIMO system with \( M > 1 \) cells (i.e., a cluster of adjacent cells in wireless cellular systems), where the same carrier frequency is used in all cells. Examples of such systems with \( M = 2 \) and \( M = 3 \) are shown in Fig. 2.1 (a) and (b), respectively. In each cell, a base station equipped with \( N_T \) transmit antennas is located at the cell center and serves a number of users, each of which is equipped with \( N_R \) receive
Chapter 2. Scheduling with Coordination for Multi-cell MIMO Networks

antennas. The base stations are connected with each other (or connected to a central controller) via backhaul links with limited capacity. We assume that the base stations can exchange control messages such as CSI and scheduling decisions but do not share user data. We consider a time slotted system, where each time frame is divided into slots of equal length. Each base station serves at most one user in a time slot. We assume a frequency flat block fading channel model, where the channel gain remains constant during a time slot and is independent and identically distributed (i.i.d.) in different time slots. This i.i.d. channel model has been widely adopted for performance analysis in the literature [60, 63, 64]. However, the proposed scheduling mechanism in this chapter can also be extended to non-i.i.d. channel models, an interesting topic for future work. We denote the set of cells as $\mathcal{I} = \{1, 2, \ldots, M\}$, and the number of users in cell $i \in \mathcal{I}$ as $K_i$. The set of users in cell $i$ is denoted as $\mathcal{K}_i$. The signal received at user $k \in \mathcal{K}_i$ can be represented as

$$y_{ik} = \sqrt{g_{ik}} H_{ik} x_i + \sum_{j \neq i, j \in \mathcal{I}} \sqrt{g_{jk}} H_{jk} x_j + n_{ik}, \quad i \in \mathcal{I},$$

(2.1)

where $x_i$ is the signal transmitted by the $i$th base station, $H_{ik} \in \mathbb{C}^{N_R \times N_T}$ is the channel matrix from the $i$th base station to user $k$, whose elements are i.i.d. and follow a complex Gaussian distribution with zero mean and unit variance ($\mathcal{CN}(0, 1)$), $g_{ik}$ is the distance-dependent average path gain from the $i$th base station to user $k$, and $n_{ik}$ is the additive white Gaussian noise (AWGN) with complex Gaussian distribution $\mathcal{CN}(0, 1)$. The first term on the right hand side in (2.1) represents the desired signal and the second term corresponds to the inter-cell interference.

Similar to some previous works (e.g., [25, 60]), we assume that mobile devices are able to perfectly estimate the desired channel and the interference channel, and this perfect CSI is available to the base station via feedback (or estimation via the uplink in time-division
duplexing systems) at the beginning of each time slot. The base stations can exchange the
CSI to make scheduling decisions. In practice, a mobile device can estimate the CSI of
each of its links based on the pilot signal received from the corresponding base stations,
i.e., we can use existing approaches such as linear minimum mean square error (LMMSE)
estimation [65] with successive interference cancellation (SIC) [66] to estimate the CSI of
each link. Since the CSI is obtained via estimation and may be erroneous, we discuss the
impact of imperfect CSI in Section 2.4.3.

2.3.2 System Stability

We consider the scenario where the data to be transmitted arrive at their associated base
station according to a stationary process. The base station maintains a transmission queue
for each of its intended users. Let $Q_{ik}[t]$ represent the queue backlog for user $k$ at the $i$th
base station at the beginning of time slot $t$. We denote the corresponding data arrival rate
and service rate for user $k$ during time slot $t$ as $A_{ik}[t]$ and $R_{ik}[t]$, respectively. Then, the
system queues evolve according to the following stochastic difference equation

$$Q_{ik}[t+1] = \max\{Q_{ik}[t] - R_{ik}[t], 0\} + A_{ik}[t], \quad \forall \ k \in \mathcal{K}_i, \ i \in \mathcal{I}.$$  \hspace{1cm} (2.2)

We define stability of the above queueing system as follows [67, p. 19].

**Definition 2.1** A discrete-time queue $Q_{ik}$ is strongly stable if

$$\limsup_{t \to \infty} \frac{1}{t} \sum_{\tau=0}^{t-1} \mathbb{E}[Q_{ik}[\tau]] < \infty.$$  \hspace{1cm} (2.3)

The system is stable if all queues in the system are strongly stable.

The above definition implies that the system is stable when the average backlog of
each queue is bounded. It has been shown in [67, p. 19] that to guarantee the stability
of the system, the average data arrival rate of each queue should be no larger than the corresponding average service rate. That is, $\overline{A}_{ik} \leq \overline{R}_{ik}, \forall k \in \mathcal{K}_i, i \in \mathcal{I}$, where

$$\overline{A}_{ik} = \lim_{t \to \infty} \frac{1}{t} \sum_{\tau=0}^{t-1} A_{ik}[\tau],$$

(2.4)

and

$$\overline{R}_{ik} = \lim_{t \to \infty} \frac{1}{t} \sum_{\tau=0}^{t-1} R_{ik}[\tau].$$

(2.5)

System stability is important since the total buffer size of a base station is finite in practice. Therefore, the downlink scheduling mechanism should guarantee that all queues in the system are strongly stable. To better assist the analysis of the scheduling mechanism, it is useful to define the stability region of the system.

**Definition 2.2** The stability region is defined as the closure of the set of all arrival rate vectors for which there exists a scheduling policy that can stabilize the system.

Let $\pi$ denote a feasible scheduling policy which results in long term average achievable data rates $\overline{R}_{ik}(\pi), \forall k \in \mathcal{K}_i, i \in \mathcal{I}$. We further denote the set of feasible scheduling policy as $\Pi$. Then, the stability region of the considered system can be represented as

$$\Lambda = \text{coh} \bigcup_{\pi \in \Pi} \{ \overline{A}_i \in \mathbb{R}^{(|\mathcal{K}_i|)} : \overline{A}_{ik} \leq \overline{R}_{ik}(\pi), \forall k \in \mathcal{K}_i, \forall i \in \mathcal{I} \},$$

(2.6)

where $|\cdot|$ is the cardinality of the set.

### 2.3.3 MIMO Transmission Strategies

To achieve high data rates in MIMO systems, we consider two physical layer transmission strategies, namely, spatial multiplexing and interference alignment, which are described below.
Chapter 2. Scheduling with Coordination for Multi-cell MIMO Networks

i) Spatial multiplexing: For a MIMO link without interference, the transmitter can deliver multiple data streams to the receiver using spatial multiplexing [68, p. 334]. For an $N_T \times N_R$ MIMO link from the $i$th base station to user $k$, $N = \min\{N_T, N_R\}$ data streams are multiplexed by using a precoding matrix $V_{ik}$ at the transmitter and are reconstructed with a decoding matrix $U_{ik}$ at the receiver. The matrices $V_{ik}$ and $U_{ik}$ are obtained from the singular value decomposition (SVD) of the channel matrix $H_{ik}$,

$$H_{ik} = U_{ik}^H \Lambda_{ik} V_{ik}, \quad (2.7)$$

where $U_{ik} \in \mathbb{C}^{N_R \times N_R}$ and $V_{ik} \in \mathbb{C}^{N_T \times N_T}$ are unitary matrices, $\Lambda_{ik} \in \mathbb{R}^{N_R \times N_T}$ is a rectangular matrix with non-negative main diagonal elements $\{\lambda_{ik,1}, \ldots, \lambda_{ik,N}\}$ and all other elements equal to zero. By applying the above precoding and decoding matrices, the MIMO link is transformed into several parallel Gaussian channels, which can support multiple data streams. The total power is distributed among the data streams using waterfilling to maximize the achievable rate (assuming there is no interference). The power allocated to the $m$th data stream for user $k$ in cell $i$ is $P_{ik,m} = \max\left\{\mu - \frac{1}{\lambda_{ik,m}}, 0\right\}$, where $\mu$ is chosen to satisfy $\sum_{m=1}^{N} P_{ik,m} = P_T$.

The achievable data rate using spatial multiplexing in the presence of inter-cell interference can be derived for the multi-cell MIMO system. In particular, the data rate for user $k$ in cell $i$ is [37]

$$R_{ik}^{SM} = \log_2 \det \left( I_{N_R} + (g_{ik} H_{ik} \Phi_{SM,i} H_{ik}^H) \left( I_{N_R} + \sum_{j \neq i} g_{jk} H_{jk} \Phi_{SM,j} H_{jk}^H \right)^{-1} \right), \quad (2.8)$$

where $I_{N_R}$ is the $N_R \times N_R$ identity matrix, $\Phi_{SM,i} = V_{ik} Q_{ik} V_{ik}^H$ is the covariance matrix of the transmitted signal at the $i$th base station, and $Q_{ik}$ is an $N_T \times N_T$ diagonal matrix with allocated power $P_{ik,m}$ ($m \in \{1, \ldots, N\}$) as main diagonal elements.
ii) **Interference alignment**: When users experience significant inter-cell interference, the performance of spatial multiplexing is limited [69]. To combat interference and improve the data rate, interference alignment can be used. It allows interference-free concurrent transmissions at the expense of sacrificing some degrees of freedom. Interference alignment requires cooperation of the base stations to jointly design the transmit and receive matrices such that the interference and the desired signal lie in orthogonal subspaces at the receiver [35]. Specifically, the $i$th base station may transmit $d_i$ ($d_i \leq N - 1$) data streams to user $k$ by designing precoding matrix $\tilde{V}_{ik}$ and decoding matrix $\tilde{U}_{ik}$ such that

$$\tilde{U}_{ik}^H H_{jk} \tilde{V}_{jk} = 0, \quad \forall \ j \in \mathcal{I}, \ j \neq i,$$

(2.9)

and

$$\text{rank} \left( \tilde{U}_{ik}^H H_{ik} \tilde{V}_{ik} \right) = d_i,$$

(2.10)

where $\tilde{U}_{ik} \in \mathbb{C}^{N_R \times d_i}$, $\tilde{V}_{ik} \in \mathbb{C}^{N_T \times d_i}$ are truncated unitary matrices, and $0$ is a $d_i \times d_i$ all-zero matrix. If we can find $\tilde{U}_{ik}$ and $\tilde{V}_{ik}, \forall \ i \in \mathcal{I}, k \in \mathcal{K}_i$ that satisfy (2.9) and (2.10), then interference can be suppressed at the desired receiver. By employing the interference alignment strategy, the achievable rate for user $k$ in cell $i$ is

$$R_{IA}^{ik} = \log_2 \det \left( I_{d_i} + g_{ik} \hat{H}_{ik} \Phi_{IA,i} \hat{H}_{ik}^H \right),$$

(2.11)

where $\hat{H}_{ik} = \tilde{U}_{ik}^H H_{ik} \tilde{V}_{ik}$, and $\Phi_{IA,i} = (P_T/d_i) I_{d_i}$ is the covariance matrix of the data symbols to be transmitted (without precoding) at the $i$th base station (with equal power allocation for all data streams). For example, in a three-cell MIMO system with $N_T = N_R = 2$, we can design $2 \times 1$ precoding vectors $(\mathbf{v}_i, i = 1, 2, 3)$ and decoding vectors $(\mathbf{u}_i, i = 1, 2, 3)$ with $d_i = 1$ such that each base station can transmit one data stream successfully to its intended receiver without any interference. Specifically, we can choose
Chapter 2. Scheduling with Coordination for Multi-cell MIMO Networks

\( \mathbf{v}_1 \) to be one of the eigenvectors of \( \mathbf{H} \mathbf{v}_1 = \mathbf{H}_{12}^{-1} \mathbf{H}_{32} \mathbf{H}_{31}^{-1} \mathbf{H}_{21} \mathbf{H}_{23}^{-1} \mathbf{H}_{13} \), and choose \( \mathbf{v}_2 = \mathbf{H}_{23}^{-1} \mathbf{H}_{13} \mathbf{v}_1 \) and \( \mathbf{v}_3 = \mathbf{H}_{31}^{-1} \mathbf{H}_{21} \mathbf{H}_{23}^{-1} \mathbf{H}_{13} \mathbf{v}_1 \). It can be shown that these precoding vectors satisfy \( \mathbf{H}_{13} \mathbf{v}_1 = \mathbf{H}_{23} \mathbf{v}_2 \), \( \mathbf{H}_{31} \mathbf{v}_3 = \mathbf{H}_{21} \mathbf{v}_2 \), and \( \mathbf{H}_{12} \mathbf{v}_1 = \mathbf{H}_{32} \mathbf{v}_3 \), which implies that the interference at each receiver can be aligned in the same subspace and the desired signal can be successfully decoded by choosing a decoding vector orthogonal to that subspace. A more detailed explanation of this design can be found in [38]. Note that in this chapter, we do not consider optimization of the power allocation for interference alignment. To simplify the analysis, we assume equal power allocation for all data streams. This assumption has also been widely used in other works (e.g. [36, 37]).

Although interference alignment can suppress the inter-cell interference, the signal power that lies in the interference subspace is lost. When the interference is not significant, spatial multiplexing may achieve a better performance [55]. Therefore, for a system with users who experience different levels of inter-cell interference, a proper transmission strategy selection is desirable for each user when scheduling the transmission.

### 2.3.4 Problem Formulation

In the above system, the downlink scheduling problem is to maximize a utility function by selecting in each cell a user to be served and a corresponding MIMO transmission strategy. The utility function is usually chosen as a concave, non-decreasing function of the service rates and should reflect a certain fairness criterion. In this chapter, we consider the proportional fair utility [70], which is a function of the long term average service rates \( \langle \bar{R}_{ik} \rangle \) of all users \( (k \in \mathcal{K}_i, i \in \mathcal{I}) \) in the system. Let vector \( \overline{\mathbf{R}} = (\bar{R}_{ik}, k \in \mathcal{K}_i, i \in \mathcal{I}) \). The
considered utility function is denoted as

$$
\phi(\mathbf{R}) = \sum_{k \in K, i \in I} \log(R_{ik}).
$$

(2.12)

Since the data rate for a user that is served by one base station depends on the interference coming from other base stations, in order to optimize the system performance, coordination among the base stations is necessary to determine the scheduling decision. To simplify the analysis, in this chapter, we propose to use the same transmission strategy at all base stations. This is reasonable since interference alignment requires all base stations to cooperate, i.e., each base station transmits only one data stream during a time slot. Nevertheless, we note that the analysis can be extended to scenarios where different transmission strategies are used by different base stations. We introduce $s[t]$ as the indicator of the transmission strategy that is selected in time slot $t$, where

$$
s[t] = \begin{cases} 
1, & \text{if spatial multiplexing is used,} \\
0, & \text{if interference alignment is used,}
\end{cases}
$$

(2.13)

and define $\mathcal{S} = \{0, 1\}$. We further introduce $l_k[t]$ to denote whether user $k$ is selected in time slot $t$, where $l_k[t] = 1$ if user $k$ is selected and $l_k[t] = 0$ otherwise. Then, the joint scheduling and transmission strategy selection problem can be formulated as

$$
\begin{align*}
\text{maximize} & \quad \phi(\mathbf{R}) \\
\text{subject to} & \quad \sum_{k \in K_i} l_k[t] = 1, \quad \forall \ i \in \mathcal{I}, \ t \in \{1, 2, \ldots\} \\
& \quad R_{ik} \geq A_{ik}, \quad \forall \ k \in \mathcal{K}_i, \ i \in \mathcal{I},
\end{align*}
$$

(2.14)

where

$$
R_{ik} = \lim_{t \to \infty} \frac{1}{t} \sum_{\tau=0}^{t-1} l_k[\tau](s[\tau]R_{ik}^{\text{SM}}[\tau] + (1 - s[\tau])R_{ik}^{\text{IA}}[\tau]).
$$

(2.15)
The last constraint in problem (2.14) implies that all queues in the system should be stable. In general, it is difficult to find an optimal solution $\mathbf{R}^*$ for this problem, since solving problem (2.14) requires CSI for all time slots, which is not possible in practice. However, it has been shown in previous work [67, p. 99] that near optimal dynamic scheduling is possible for problems with a structure similar to that of (2.14) by using a stochastic network optimization approach, which will be discussed in the following section.

2.4 Dynamic Scheduling Mechanisms

In this section, we develop a dynamic scheduling framework with transmission strategy selection based on a stochastic network optimization approach, and propose centralized and distributed scheduling mechanisms under perfect CSI. We also discuss the impact of imperfect CSI at the base stations and propose a rate adjustment scheme to improve the performance.

2.4.1 Centralized Dynamic Scheduling Mechanism

We first consider the scenario where global CSI and queue backlog information are available at the scheduler (either a base station or a central controller). Note that problem (2.14) involves optimizing a concave function of time average rates by making scheduling decisions in each time slot, which has a similar structure to the problem discussed in [67, p. 99]. Therefore, we can apply a stochastic network optimization approach to the system considered in Section 2.3. The main idea to solve problem (2.14) is to use a weighted sum-rate maximization algorithm which operates in each time slot with causal CSI. The weight of each user’s rate is updated according to the user’s actual queue backlog and the virtual queue backlog of an auxiliary variable in order to satisfy the stability and fairness criteria. Based on this approach, we obtain a dynamic
Chapter 2. Scheduling with Coordination for Multi-cell MIMO Networks

scheduling framework with transmission strategy selection as follows. We first introduce
an auxiliary variable $\gamma_{ik}[t]$ for each user $k$ in cell $i$ at time slot $t$, and let $W_{ik}[t]$ represent
the virtual queue backlog associated with $\gamma_{ik}[t]$. The virtual queues evolve according to
the following stochastic difference equation

$$W_{ik}[t + 1] = \max\{W_{ik}[t] - R_{ik}[t], 0\} + \gamma_{ik}[t], \quad \forall k \in K_i, \ i \in I.$$  (2.16)

Then, at the beginning of time slot $t$ ($t = 0, 1, \ldots$), the proposed dynamic scheduling
framework involves the following steps.

(i) The first step is to determine the values of auxiliary variables $\gamma_{ik}[t]$ ($k \in K_i, i \in I$) by
solving the following problem:

$$\max_{\gamma[t], i \in I} \beta \phi(\gamma[t]) - \sum_{i \in I} \sum_{k \in K_i} \gamma_{ik}[t]W_{ik}[t]$$

subject to $0 \leq \gamma_{ik}[t] \leq \gamma_{\text{max}}, \ \forall k \in K_i, \ i \in I,$

(2.17)

where $\beta, \gamma_{\text{max}} > 0$ are predetermined system parameters, $W_{ik}[t]$ is the corresponding virtual
queue backlog known at the scheduler, and vector $\gamma[t] = (\gamma_{ik}[t], k \in K_i, i \in I)$.

(ii) The next step is to apply a weighted sum rate maximization scheduling policy to the
system with actual and virtual queues. Specifically, given the queue backlog information
$Q_{ik}[t]$ and $W_{ik}[t]$ for all users, we select the users $k_i[t] \in K_i, \forall i \in I$ and the corresponding
transmission strategy $s[t] \in S$ by solving the following optimization problem

$$\max_{k_i[t], s[t]} \sum_{i \in I} (Q_{ik_i[t]}[t] + W_{ik_i[t]}[t])R_{ik_i[t]}[t].$$  (2.18)

(iii) Finally, we update all the actual queues $Q_{ik}[t + 1]$ and virtual queues $W_{ik}[t + 1]$ according to (2.2) and (2.16), respectively.

By applying the above scheduling framework, the following results can be obtained.
Proposition 2.1 Assume the data arrival rate and the service rate are upper bounded by \( A_{\text{max}} \) and \( R_{\text{max}} \), respectively. For given constants \( \beta, \gamma_{\text{max}} \), and a concave and entry-wise non-decreasing utility function \( \phi(\cdot) \), if there exists at least one feasible scheduling policy, then we have

\[
\lim_{t \to \infty} \inf \phi \left( \frac{1}{t} \sum_{\tau=0}^{t-1} \mathbb{E}\{R[\tau]\} \right) \geq \phi(\mathbf{R}^*(\gamma_{\text{max}})) - \frac{D}{\beta},
\]

(2.19)

where \( D \) is a constant that satisfies

\[
D \geq \mathbb{E} \left\{ \sum_{i \in I} \sum_{k \in K_i} \frac{1}{2} (A_{ik}[t]^2 + \gamma_{ik}[t]^2 + 2R_{ik}[t]^2) \right\}.
\]

\( \mathbf{R}[\tau] \) is the vector containing the service rates for all users in time slot \( \tau \), the expectation \( \mathbb{E}[\cdot] \) is with respect to the joint probability distribution of the channel matrix and scheduling decisions under the proposed scheduling mechanism, and \( \mathbf{R}^*(\gamma_{\text{max}}) \) is the solution to problem (2.14) with the additional constraint \( 0 \leq R_{ik} \leq \gamma_{\text{max}}, \forall k \in K_i, i \in I \).

If there is an \( \epsilon \geq 0 \) and a feasible scheduling policy \( \nu \) that gives rates \( \mathbf{R}^\nu[t] = \{R^\nu_{ik}[t], \forall k \in K_i, i \in I\} \) which satisfy \( \mathbb{E}\{A_{ik}[t] - R^\nu_{ik}[t]\} \leq -\epsilon, 0 \leq \mathbb{E}\{R^\nu_{ik}[t]\} \leq \gamma_{\text{max}}, \forall k \in K_i, i \in I \) and \( \phi(\mathbb{E}\{\mathbf{R}^\nu[t]\}) = \phi_\epsilon \), then we have

\[
\lim_{t \to \infty} \sup \frac{1}{t} \sum_{\tau=0}^{t-1} \sum_{i \in I} \sum_{k \in K_i} \mathbb{E}\{Q_{ik}[\tau]\} \leq \frac{D + \beta[\phi(\mathbf{R}^*(\gamma_{\text{max}})) - \phi_\epsilon]}{\epsilon}.
\]

(2.20)

The proof of Proposition 2.1 follows [67, Chapter 5], and a sketch of the proof is provided in Appendix A. The result in (2.19) implies that, when \( \gamma_{\text{max}} \) is sufficiently large (such that \( \mathbf{R}^*(\gamma_{\text{max}}) = \mathbf{R}^* \)), the proposed dynamic scheduling framework can achieve a utility that is arbitrarily close to the optimal value \( \phi(\mathbf{R}^*) \) by increasing \( \beta \). As a tradeoff, the actual queue backlog of the system grows linearly with \( \beta \), which can be seen from (2.20). Based on the
Chapter 2. Scheduling with Coordination for Multi-cell MIMO Networks

Algorithm 2.1 Centralized dynamic scheduling algorithm.

1: **Initialization**
2: Initialize $\beta$, $\gamma_{\text{max}}$, and $W_{ik}[0]$, $\forall \, k \in \mathcal{K}_i$, $i \in \mathcal{I}$.
3: repeat
4:  if $t \in \{0, 1, 2, \ldots\}$ then
5:    Each BS collects the CSI from its subscribed users.
6:    The scheduler obtains global CSI and queue backlog information via backhaul links.
7:    The scheduler finds $\gamma_{ik}[t]$, $\forall \, k \in \mathcal{K}_i$, $i \in \mathcal{I}$ by solving (2.17).
8:    The scheduler finds $k_i^*[t] \in \mathcal{K}_i$, $i \in \mathcal{I}$ and $s^*[t]$ by solving (2.18).
9:    The scheduler updates $W_{ik}[t + 1]$, $\forall \, k \in \mathcal{K}_i$, $i \in \mathcal{I}$ according to (2.16).
10:   The scheduler sends the scheduling decision to each base station.
11:   Each BS updates $Q_{ik}[t + 1]$, $\forall \, k \in \mathcal{K}_i$, $i \in \mathcal{I}$ according to (2.2) at the end of the time slot.
12: endif
13: until system stops operation.

above framework, we propose a centralized scheduling mechanism as shown in Algorithm 2.1.

In Algorithm 2.1, the scheduler first initializes all system parameters before the scheduling process begins. The scheduler can be either one of the base stations or a central controller that connects to all base stations. Then, at the beginning of each time slot, each base station collects the CSI of all direct links (from the base station to its subscribed users) and interference links (from other base stations to these users), and passes the CSI and queue backlog information of its subscribed users to the scheduler. Next, the scheduler numerically computes the values of auxiliary variables according to (2.17) in Line 7 and selects the desired users and the corresponding transmission strategy by searching for the optimal solution to (2.18) in Line 8. Then, the scheduler updates the virtual queue backlogs and sends the scheduling decision to each base station in Lines 9 and 10, respectively. Finally, at the end of the time slot, both base stations update their actual queues according to (2.2). This scheduling process will be repeated until the end of the transmission session.
2.4.2 Distributed Dynamic Scheduling Mechanism

The centralized dynamic scheduling mechanism requires global CSI and queue backlog information at the scheduler to make scheduling decisions. However, in practical systems, the backhaul links are capacity-limited and exchanging extensive amounts of information may incur a large communication overhead. For instance, in the considered system, the \(i\)th base station needs to send information regarding \(MK_i\) channel matrices and \(K_i\) queue backlogs to the scheduler in each time slot. Therefore, it is attractive to develop distributed scheduling mechanisms. In this section, we propose an efficient distributed dynamic scheduling mechanism which only requires limited message exchanges between the base stations. Note that designing precoding and decoding matrices for interference alignment becomes more complicated as the number of cells increases. To simplify the analysis, we focus on the design of a distributed scheduling mechanism for a two-cell system.

From the discussion in Section 2.4.1, it can be seen that the desired distributed scheduling mechanism requires distributed solutions to (2.17) and (2.18) that can be found at each base station. We first consider solving problem (2.17) distributively. Intuitively, if the utility function \(\phi(\cdot)\) can be decomposed at each base station independently, then problem (2.17) can be solved distributively. In this chapter, since we adopt the proportional fair utility \(\phi(\gamma) = \sum_{k \in K_1} \log(\gamma_{1k}) + \sum_{k' \in K_2} \log(\gamma_{2k'})\) for a two-cell system, a distributed solution for (2.17) can be obtained by solving the following optimization problem at the \(i\)th base station (for simplicity, we omit the time index \([t]\) in this subsection):

\[
\begin{align*}
\text{maximize} & \quad \sum_{k \in K_i} (\beta \log(\gamma_{ik}) - W_{ik} \gamma_{ik}) \\
\text{subject to} & \quad 0 \leq \gamma_{ik} \leq \gamma_{\max}, \quad \forall \ k \in K_i.
\end{align*}
\]  

(2.21)

The optimal solution of problem (2.21) is given by \(\gamma_{ik} = \min\{\beta/W_{ik}, \gamma_{\max}\}, \forall k \in K_i, i \in \mathcal{I}\).
Next, we develop a distributed algorithm to solve problem (2.18). The major challenge to solve the problem distributively is that the objective function is not decomposable, since the achievable rate of a user in one cell depends on the scheduling decision in the other cell. Specifically, when calculating the achievable rate for spatial multiplexing independently at a base station, the covariance matrix of the interference in (2.8) is not known. Moreover, the rate in (2.9) can only be achieved when both base stations choose to use interference alignment, and therefore, coordination is needed when determining the transmission strategy. To address the above challenges, we propose a two-step approach. In the first step, we relax the constraints in (2.18) and find distributed solutions to the resulting subproblems with a fixed transmission strategy. In the second step, we design a coordination mechanism such that the base stations make final scheduling decisions with limited message exchange based on the results obtained in the first step.

**Step 1:** We introduce $\varphi_i(k_i, s) = (Q_{ik_i} + W_{ik_i})R_{ik_i}^s$, where $R_{ik_i}^s$ represents the transmission rate at the $i$th base station for user $k_i$ with strategy $s$. Since $s$ only takes binary values from $S$, for a fixed $s$, (2.18) is reduced to the following optimization problem:

$$\max_{k_1 \in K_1, k_2 \in K_2} \varphi_1(k_1, s) + \varphi_2(k_2, s)$$

To solve (2.18) distributively, we first find the distributed solution for (2.22) for different values of $s$.

**Case 1:** When $s = 0$, both base stations employ the spatial multiplexing transmission strategy. Note that each base station only has local CSI (including the CSI of the desired links and the CSI of the links between the intended users and the interfering base stations) obtained via feedback from its intended users. To solve (2.22) distributively at each base station, we adopt the following approximation.
achievable rate for the users in cell $i$ using (2.8), we use the average value of the interference covariance matrix $\mathbb{E}[\Phi_{SM,j}]$ instead of the instantaneous value $\Phi_{SM,j}$ since the scheduling decision of the other cell is not known. Based on this approximation, the objective function in (2.22) can be decomposed into two independent functions with respect to different base stations, and the distributed solution is to let base station $i$ find the user $k_i^0$ that satisfies $\varphi_i(k_i^0, 0) = \max_{k_i \in K_i} \varphi_i(k_i, 0)$.

Case 2: When $s = 1$, the base stations use the interference alignment transmission strategy simultaneously. In this chapter, for the two-cell case, we use a fixed precoding vector with equal power allocation at the base stations for interference alignment. Then, the inter-cell interference can be canceled at the users by designing decoding matrices with local CSI, and the achievable data rate can be calculated independently at each base station. Therefore, problem (2.22) can also be solved distributively by letting base station $i$ select the user $k_i^1$ that satisfies $\varphi_i(k_i^1, 1) = \max_{k_i \in K_i} \varphi_i(k_i, 1)$.

Step 2: After solving the relaxed optimization problem (2.22) in Step 1, each base station obtains two scheduling decisions with respect to different transmission strategies, i.e., the decisions at the $i$th base station are $(k_i^s, s) \forall s \in S$. The next step is to design an efficient coordination mechanism which finds the optimal scheduling decision that maximizes the overall system utility. Based on the distributed solution in Step 1, it can be easily verified that the optimal value of the objective function in (2.18) is

$$\max \{ \sum_{i \in I} \varphi_i(k_i^0, 0), \sum_{i \in I} \varphi_i(k_i^1, 1) \}.$$ 

We introduce a coordination variable $\delta_i$ for the $i$th base station, where $\delta_i = \varphi_i(k_i^0, 0) - \varphi_i(k_i^1, 1)$. Then, after exchanging this variable with the other base station, the $i$th base station obtains

$$\sum_{i \in I} \delta_i = \sum_{i \in I} \varphi_i(k_i^0, 0) - \sum_{i \in I} \varphi_i(k_i^1, 1). \quad (2.23)$$

It can be seen that when $\sum_{i \in I} \delta_i > 0$, the scheduling decision with spatial multiplexing is
desired, otherwise the scheduling decision with interference alignment is preferable. Therefore, the base stations can find their scheduling decisions by exchanging only the value of the coordination variable $\delta_i$, and the optimal decision for the $i$th base station is given by

$$ (k_i^*, s^*) = \begin{cases} 
(k_i^0, 0), & \text{if } \sum_{i \in I} \delta_i > 0, \\
(k_i^1, 1), & \text{if } \sum_{i \in I} \delta_i \leq 0.
\end{cases} \quad (2.24) $$

Based on the above discussion, we propose the distributed scheduling mechanism described in Algorithm 2.2. Different from Algorithm 2.1, in Line 7 of Algorithm 2.2, the base stations find their potential scheduling decisions based on local CSI and queue backlog information using the proposed two-step approach. Then, they compute and exchange the value of the coordination variable in Line 8, and make the final scheduling decisions according to (2.24) in Line 9. The rest of the algorithm is similar to Algorithm 2.1. In Algorithm 2.2, the scheduling problem is solved distributively at each base station with limited information exchange (only the value of one variable is exchanged). Therefore, the communication overhead is reduced, with the trade-off that the scheduling decisions are suboptimal since an approximation is used when solving problem (2.22). Although Algorithm 2.2 is designed for a two-cell system, it can be extended to systems with more cells with proper adjustment. For example, for a three-cell system, when solving the distributed scheduling problem with $s = 1$ in Line 7 of Algorithm 2.2, we may use average data rate under different channel realizations as an approximation of the instantaneous data rate for each user, i.e., we simulate different global CSI at a base station and numerically compute the average data rate for its intended user assuming centralized design for interference alignment (as shown in Section 2.3.3). Other steps remain the same as those in Algorithm 2.2. Note that the objective of the distributed scheduling algorithm is to determine the desired user to be served and the corresponding transmission strategy for each base station. Once these
Chapter 2. Scheduling with Coordination for Multi-cell MIMO Networks

Algorithm 2.2 Distributed scheduling algorithm executed at the $i$th base station.

1: **Initialization**
2: Initialize $\beta$, $\gamma_{\text{max}}$, and $W_{ik}[0]$, $\forall k \in \mathcal{K}_i$.
3: **repeat**
4: if $t \in \{0, 1, 2, \ldots\}$ then
5: Obtain all the CSI from subscribed users.
6: Find $\gamma_{ik}[t]$, $\forall k \in \mathcal{K}_i$ by solving (2.17).
7: Find $(k_{i0}^0[t], \varphi_i(k_{i0}^0[t], 0))$ and $(k_{i1}^1[t], \varphi_i(k_{i1}^1, 1))$ by solving (2.22) with $s = 0$ and $s = 1$.
8: Compute $\delta_i[t]$ and exchange it with the other base station to obtain $\sum_{i \in \mathcal{I}} \delta_i[t]$.
9: Obtain the scheduling decision according to (2.24).
10: Update $W_{ik}[t + 1]$, $\forall k \in \mathcal{K}_i$ according to (2.16).
11: Update $Q_{ik}[t + 1]$, $\forall k \in \mathcal{K}_i$ according to (2.2) at the end of the time slot.
12: **endif**
13: **until** system stops operation.

decisions are made, base stations can cooperate with each other to transmit their data, i.e., interference alignment with local CSI can be achieved by using existing distributed algorithms such as those in [71] and [72].

2.4.3 Scheduling with Imperfect CSI

The dynamic scheduling mechanisms proposed in the previous subsections assume perfect CSI at the base stations. However, in practice, the channel matrices at the base stations may not be perfect (due to delayed or erroneous feedback from the user terminals). Therefore, it is of great importance to analyze the impact of imperfect CSI on the performance of the proposed scheduling mechanisms. We adopt the CSI model for MIMO systems from [73] and [74], where we assume the CSI is perfect at the user terminals but is imperfect at the base stations. This model is realistic as the user terminal can update its channel estimate frequently exploiting the pilots sent by the base stations. On the other hand, the channel estimate at the base station may be updated less frequently to reduce the feedback and computational cost. The imperfect CSI at the $i$th base station (for user $k$) is modeled as $\hat{H}_{ik} = H_{ik} + e\Psi$, where $H_{ik}$ is the true channel matrix and $e\Psi$
is the error incurred during feedback which is statistically independent from $H_{ik}$. The elements of $\Psi$ are i.i.d. zero mean complex Gaussian variables with unit variance, and $e$ (where $0 \leq e \leq 1$) is a scalar that characterizes how accurate the feedback is. In this scenario, since the precoding and decoding matrices are designed at the scheduler (or the base stations) based on the imperfect CSI, the achievable rates obtained from (2.8) and (2.11) are inaccurate. For user $k$ in cell $i$, we denote the precoding and decoding matrices with spatial multiplexing under imperfect CSI as $\hat{V}_{ik}$ and $\hat{U}_{ik}$, respectively, and the corresponding precoding and decoding matrices with interference alignment as $\tilde{V}_{ik}$ and $\tilde{U}_{ik}$, respectively. We further denote $\hat{R}_{is}^s$ as the corresponding estimated data rate under transmission strategy $s$, which is calculated according to (2.8) with $\hat{V}_{ik}$ and $\hat{U}_{ik}$ (or (2.11) with $\tilde{V}_{ik}$ and $\tilde{U}_{ik}$), and let $\hat{R}_{is}^{*\dagger}$ be the corresponding actual achievable rate. Then, we have

$$
\hat{R}_{is}^s = \begin{cases} 
\log_2 \det \left( I_{N_R} + (g_{ik} \hat{H}_{ik} \hat{\Phi}_{SM,i} \hat{H}_{ik}^H) \times (I_{N_R} + \sum_{j \neq i} g_{jk} \hat{H}_{jk} \hat{\Phi}_{SM,j} \hat{H}_{jk}^H)^{-1} \right), & s = 0, \\
\log_2 \det \left( I_{d_i} + g_{ik} \hat{H}_{ik}^\dagger \hat{\Phi}_{IA,i} (\hat{H}_{ik}^\dagger)^H \right), & s = 1,
\end{cases} \tag{2.25}
$$

and

$$
\hat{R}_{is}^{*\dagger} = \begin{cases} 
\log_2 \det \left( I_{N_R} + (g_{ik} \hat{H}_{ik} \hat{\Phi}_{SM,i} \hat{H}_{ik}^H) \times (I_{N_R} + \sum_{j \neq i} g_{jk} \hat{H}_{jk} \hat{\Phi}_{SM,j} \hat{H}_{jk}^H)^{-1} \right), & s = 0, \\
\log_2 \det \left( I_{d_i} + g_{ik} \hat{H}_{ik}^{*\dagger} \hat{\Phi}_{IA,i} (\hat{H}_{ik}^{*\dagger})^H \times (I_{d_i} + \sum_{j \neq i} g_{jk} \hat{H}_{jk}^{\dagger} \hat{\Phi}_{IA,j} (\hat{H}_{jk}^{\dagger})^H)^{-1} \right), & s = 1,
\end{cases} \tag{2.26}
$$

where $\hat{\Phi}_{SM,i}$ and $\hat{\Phi}_{IA,i}$ are transmit covariance matrices designed according to the imperfect channel, $H_{ik}' = \tilde{U}_{ik} \hat{H}_{ik} \tilde{V}_{ik}$, and $H_{ik}^\dagger = \hat{U}_{ik} \hat{H}_{ik} \hat{V}_{ik}$. According to Shannon’s channel coding, when $\hat{R}_{is}^s > \hat{R}_{is}^{*\dagger}$, the receiver cannot successfully decode the data and an outage
event occurs. Therefore, the effective data rate with imperfect CSI for any \( s \in S \) is

\[
R_{ik}^{*s} = \begin{cases} 
\hat{R}_{ik}^s, & \text{if } \hat{R}_{ik}^s \leq \hat{R}_{ik}^{s\dagger}, \\
0, & \text{otherwise}.
\end{cases}
\] (2.27)

We denote the corresponding transmission outage probability as \( \xi_{ik}^s = \Pr\{\hat{R}_{ik}^s > \hat{R}_{ik}^{s\dagger}\} \).

Intuitively, when \( \xi_{ik}^s \) is large, scheduling based on \( \hat{R}_{ik}^s \) is inefficient and the system performance may degrade severely. To address this problem, we extend the proposed scheduling mechanisms by introducing a weighting factor to adjust the transmission rate for each user. Specifically, when searching for the optimal scheduling decision, we use \( p_{ik}^s \hat{R}_{ik}^s \) instead of \( \hat{R}_{ik}^s \) as the transmission rate, where \( p_{ik}^s > 0 \) is a weighting factor for user \( k \) in cell \( i \) when transmission strategy \( s \) is employed. We find \( p_{ik}^s \) by running simulations over different channel realizations, and select the value such that the average outage probability with adjusted service rate is smaller than a certain value \( \xi_0 \). For example, for a particular user \( k \), we search by gradually increasing \( p_{ik}^s \) from 0 with a step size of \( \Delta p \).

For each \( p_{ik}^s \), we simulate \( N_s \) channel realizations, and count the total number of outage events \( N_{out} \) (where \( p_{ik}^s \hat{R}_{ik}^s > \hat{R}_{ik}^{s\dagger} \)). We select the largest value of \( p_{ik}^s \) such that \( N_{out}/N_s < \xi_0 \). By adjusting the data rate with the weighting factor, the system performance is improved since the average outage probability for a user is limited to a small value and the transmission becomes more efficient.

## 2.5 Performance Evaluation

We evaluate the performance of the proposed dynamic scheduling mechanisms via simulations. We first consider a two-cell MIMO system, where the radius of each cell is 500 m and the distance between the two base stations is \( d \). In each cell, 10 users are
randomly deployed and at least \( N_e \) of them are located in the cell-edge region. For any position in the cell-edge region of the \( i \)th cell, the distance from this position to base station \( i \) is between 450 m and 500 m, and the angle between the direction from base station \( i \) to this position and the direction from base station \( i \) to the other base station is no more than 30°. The base stations and the users are equipped with two omni-directional antennas. All wireless links experience Rayleigh fading and path loss, and the 3GPP urban-micro path loss model is used [75]. The transmission power at a base station is \( P_T \). The common channel has a bandwidth of 200 kHz with noise variance \(-120 \text{ dBm}\). The duration of a time slot is 100 ms. The data packet size is 8 kbits. We consider Poisson arrival processes with different average data rates for users, where the average arrival data rate for a user is randomly selected from the set \( \{3, 4, 5, 6\} \) (packets per time slot). The base stations store 50 packets for each user before transmission starts. Other parameters are \( R_{\text{max}} = \gamma_{\text{max}} = 100 \text{ bit/s/Hz}, \beta = 1000 \).

We first evaluate the performance of the proposed scheduling mechanisms for perfect CSI. We implement the proposed centralized scheduling mechanism according to
Algorithm 2.1. We also implement scheduling mechanisms which use a single transmission strategy with the proposed stochastic network optimization approach. Fig. 2.2 shows the average utility obtained by different scheduling mechanisms under varying transmission power \( P_T \), where \( d = 1 \) km and \( N_e = 3 \) cell edge users. It is shown that the average utility of all the scheduling mechanisms increase as \( P_T \) increases, and the proposed scheduling mechanism outperforms the mechanisms with a single transmission strategy. As \( P_T \) decreases, the performance of the proposed scheduling mechanism approaches that of the mechanism with spatial multiplexing. The reason is that, when the transmission power decreases, the interference power becomes insignificant and noise dominates the performance. Therefore, the spatial multiplexing transmission strategy is preferable for almost all the users. On the other hand, as the transmission power increases, the interference power becomes larger while the noise power becomes comparably insignificant, and eliminating interference becomes critical. Hence, interference alignment has superior performance for almost all the users, and the performance of the proposed scheduling mechanism approaches that of the mechanism with interference alignment.

Fig. 2.3 shows the average utility for different distances between the base stations \( (d) \), where \( P_T = 20 \) dBm and \( N_e = 3 \). It can be seen that the average utility of the scheduling mechanism using interference alignment remains almost constant, while the average utility of the other scheduling mechanisms increase as \( d \) increases. In addition, the performance of the proposed scheduling mechanism is close to that of the mechanism with interference alignment when \( d \) is small and it approaches that of the mechanism with spatial multiplexing as \( d \) increases. The reasons for this behavior are as follows. First, since the transmission power is fixed, the inter-cell interference becomes stronger as the base stations get closer, which degrades the performance of spatial multiplexing.
Chapter 2. Scheduling with Coordination for Multi-cell MIMO Networks

Figure 2.3: Two-cell model: the average utility versus the distance $d$ for $P_T = 20$ dBm and $N_e = 3$.

However, since the performance of interference alignment only depends on the transmission power of the desired signal (but not on the interference power), the average utility of the scheduling scheme with interference alignment remains almost constant. Second, when the base stations are close to each other ($d$ is small), the inter-cell interference dominates the performance, and the interference alignment strategy is preferable for most of the users. On the contrary, when the two base stations are far apart from each other ($d$ is large), the impact of the inter-cell interference becomes insignificant compared to the noise, and spatial multiplexing becomes superior for most of the users in the system.

The performance of the proposed scheduling mechanism for different numbers of cell-edge users ($N_e$) is shown in Fig. 2.4, where $P_T = 20$ dBm and $d = 1$ km. As can be observed, the performance of all scheduling mechanisms degrades as $N_e$ increases. Specifically, the performance of the proposed scheduling mechanism is close to that of the mechanism with spatial multiplexing when $N_e$ is small, while it approaches the performance of the mechanism with interference alignment when $N_e$ is large. This is because on average interference alignment outperforms spatial multiplexing for users in the cell-edge region.
Therefore, when there are few cell-edge users, the proposed scheduling mechanism will choose to use spatial multiplexing most of the time. On the other hand, when the system is dominated by cell-edge users, the proposed scheduling mechanism tends to use interference alignment.

Next, we compare the performance of the proposed scheduling mechanism with a weighted sum-rate maximization (WSRM) scheduling mechanism. The WSRM scheduling aims to maximize the weighted sum of all users’ rates in each time slot, where the weight for a user is chosen as its corresponding queue backlog. It is shown in [76] that this scheduling mechanism guarantees the system stability when the average arrival data rates are in the feasible region. In Fig. 2.5, we show the average utility for the proposed centralized and distributed scheduling mechanisms, as well as the WSRM mechanism for which the transmission strategy selection was implemented in a centralized manner. As can be seen, the proposed distributed scheduling mechanism achieves almost the same performance as the centralized scheduling mechanism, which demonstrates its effectiveness. Both the proposed mechanisms achieve a higher utility than the WSRM
mechanism, which implies that the proposed mechanisms are superior in fairly allocating the resources to the users while stabilizing the system.

We also evaluate the performance of the proposed scheduling mechanisms for imperfect CSI. We adopt the imperfect CSI model in Section 2.4.3, and set $e = 0.05$. The other simulation parameters are identical to those in Fig. 2.2. Fig. 2.6 shows the average utility of the proposed mechanisms with and without rate adjustment for different transmission powers ($P_T$), where we set $\xi_0 = 0.05$. The scheduling mechanisms with rate adjustment achieve significant performance improvement compared to those without rate adjustment. Fig. 2.7 shows the average utility of the proposed mechanisms with rate adjustment for different $\xi_0$ values. It is shown that the average utility first increases with $\xi_0$ and then decreases when $\xi_0$ exceeds some value. The reason for this behavior is as follows. According to Section 2.4.3, a larger $\xi_0$ implies a larger weighting factor ($p_{ik}^s$), which tends to increase the effective service rate. Increasing $\xi_0$ may also increase the outage probability, which tends to degrade the effective service rate. When $\xi_0$ is small, increasing the weighting factor has a greater impact and the average effective service rate becomes larger. However,
Figure 2.6: Two-cell model: the average utility versus transmission power $P_T$ for $\xi_0 = 0.05$.

Figure 2.7: Two-cell model: the average utility versus $\xi_0$ for $P_T = 20, 30$ dBm.
Figure 2.8: Three-cell model: the average utility versus transmission power $P_T$ for perfect CSI.

when $\xi_0$ exceeds some values, the impact of increasing the outage probability becomes significant and the average effective service rate starts to decrease.

Finally, we evaluate the performance of the proposed scheduling mechanisms for a three-cell model. The system model is shown in Fig. 2.1 (b), where the distance between any two base stations is 1 km and 10 users are randomly deployed in each cell. Other system parameters are identical to those of the two-cell model. The three-cell interference alignment scheme is implemented according to [9]. We evaluate the performance for both perfect and imperfect CSI. Fig. 2.8 shows the performance for different transmission powers at the base stations, where the proposed scheduling mechanism achieves a better performance compared to mechanisms without transmission strategy selection. 2.9 shows the results for an imperfect CSI scenario with $e = 0.03$ and $\xi_0 = 0.05$, where the proposed centralized and distributed mechanisms with rate adjustment achieve superior performance compared to those without rate adjustment. Note that the results in Figs. 2.8 and 2.9 are consistent with those for the two-cell model, which demonstrates the effectiveness of the proposed strategy.
In this chapter, a downlink scheduling mechanism with transmission strategy selection was proposed for multi-cell MIMO networks. The proposed mechanism employed two MIMO transmission strategies, spatial multiplexing and interference alignment, to improve the system performance. Based on a stochastic optimization technique, a centralized dynamic scheduling algorithm was proposed which jointly selects a user and the corresponding transmission strategy for each base station to maximize the overall system utility while keeping the system stable. Then, a distributed scheduling algorithm was developed which only requires limited message exchanges between the base stations. Furthermore, a rate adjustment approach is introduced to improve the performance of the proposed scheduling mechanism when the CSI is imperfect. Simulation results showed that the performance of the proposed distributed scheduling mechanism is close to that of the centralized scheduling mechanism, and both mechanisms achieved a better performance than mechanisms with a single transmission strategy, especially for the case where inter-cell interference dominates the performance of some users.
Chapter 3

Network Configuration for Two-tier Heterogeneous Networks

3.1 Introduction

Recent attention on two-tier macro-femto networks has been drawn to designing efficient network configuration mechanisms (i.e., spectrum allocation, power management) with access control at the FBSs [41–48, 77–81]. In this chapter, uplink network configuration is studied for a two-tier macro-femto network, where an MBS and a cluster of adjacent FBSs together serve users with different QoS requirements. Different from existing works, in this chapter, base stations and users are considered as individual decision makers with different objectives. As discussed in Chapter 1, in two-tier macro-femto networks, employing hybrid access at the FBSs can improve the overall system performance, especially when MUs experience better signal quality from the FBS than from the MBS. However, incentive should be provided for FBSs to operate at hybrid access mode since they are installed by private FUs who may not share their resources to MUs for free. Such incentive mechanisms are typically designed based on certain business models and involve trading mechanisms, which may not be available for the current system. In this chapter, a dynamic resource allocation approach is employed to provide incentive for hybrid access at FBSs. In this approach, the number of channels allocated to an FBS depends on its access control decisions, which motivates the FBS to adopt hybrid access.
Chapter 3. Network Configuration for Two-tier Heterogeneous Networks

if more channels are desirable. In addition to providing incentives for using hybrid access at the FBSs, this chapter aims to design an efficient network configuration mechanism which coordinates the access control, resource allocation and power management processes. To achieve this, the network configuration problem is modeled as a five-level optimization problem, and is analyzed using a bottom-up approach. Efficient algorithms are proposed to obtain the optimal decision for each network configuration process.

The rest of this chapter is organized as follows. Section 3.2 provides a summary of related works. In Section 3.3, we describe the system model and network configuration mechanism, and formulate the multi-level optimization problem. In Section 3.4, we analyze the multi-level optimization problem and design the network configuration mechanism based on the algorithm developed in each optimization level. Simulation results are presented in Section 3.5, and summary is given in Section 3.6.

3.2 Related Works

Incentive mechanisms have been proposed in the literature to motivate hybrid access in two-tier macro-femto networks [44–48]. In [44], a spectrum leasing mechanism has been proposed, where the MBS leases the spectrum to the FBSs, and the FBSs further lease their spectrum to nearby MUs. The mechanism is formulated as a Stackelberg game and the optimal spectrum leasing prices are determined via equilibrium analysis. Similar approaches have been employed in [45] and [46] with different optimization objectives. An access permission based incentive mechanism has been proposed in [47], where the macrocell service provider purchases access permissions from multiple competing femtocell service providers. In [48], the authors propose a spectrum-sharing rewarding framework to encourage hybrid access at FBSs, where femtocell owners determine the amount of resources shared with public users and the operator determines the ratio of revenue distribution.
to femtocell owners. The aforementioned incentive mechanisms are developed based on specific business models, which may not be available in the existing systems. Therefore, it is interesting to design resource allocation mechanisms to provide incentive without bothering the business model.

Several access control schemes have been proposed for two-tier macro-femto networks recently. In [77], FBSs grant access to the MUs who generate significant interference to the FUs in the uplink, so that the total number of interferers is reduced below a certain threshold. In [78], an MU is selected to be associated with its geographically closest FBS if the distance to this FBS is smaller than a threshold. However, these schemes are either heuristic or designed independently from other network configuration processes. In [79], joint admission control and resource allocation has been studied using the theory of semi-Markov decision process, and a distributed power adaptation algorithm is proposed to reduce energy consumption at femtocells. In [80], the authors propose optimal access control, subcarriers allocation and power control algorithms in an OFDM macrocell/femtocell network assuming hybrid access mode at the FBSs. The authors of [81] propose an on-demand waterfilling type channel allocation scheme with the consideration of user grouping for MBS and FBSs. However, the incentive for private owners to use hybrid access is not discussed in the aforementioned works.

3.3 System Model

We consider the uplink of a two-tier macro-femto wireless system, where an MBS and a cluster of \( J > 1 \) FBSs (which are close to each other) together serve a number of mobile users as shown in Fig. 3.1. This model represents the cellular network in a residential area (e.g., in a building) where FBSs are installed by home users to improve the service quality. In such networks, each FBS has a dedicated group of subscribed FUs, and the MBS has
a number of subscribed MUs that are also close to the FBSs. The FBSs connect to the MBS and the operator’s core network via wired backhaul links. We assume the capacity of the backhaul link for each FBS is large enough to handle the data transmission for its associated users. A central scheduler located at the operator’s core network is responsible for allocating the available spectrum resources to all the base stations. We denote the set of base stations as $\mathcal{I} = \{0, 1, \ldots, J\}$, where 0 indicates the MBS. For base station $j \in \mathcal{I}$, we define the set of associated MUs as $S^m_j$, and the set of associated FUs as $S^f_j$. Thus, the set of all users associated with base station $j$ is denoted as $S_j = S^m_j \cup S^f_j$. The set of all MUs is denoted as $S^M = \bigcup_{j \in \mathcal{I}} S^m_j$. A list of key notations is provided in Table 3.1.

We consider a time slotted system, where the time frame is divided into slots of equal length. We denote the set of available channels as $\mathcal{N} = \{1, 2, \ldots, N\}$, where each channel has a bandwidth of $B$. We assume a frequency flat block fading wireless channel model [82], where for any of the $N$ channels, the channel fading between a user and its base station remains constant during a time slot and is i.i.d. in different time slots with zero mean and unit variance complex Gaussian distribution.
Table 3.1: List of key notations used in this chapter.

<table>
<thead>
<tr>
<th>Notation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>$a_{ij}$</td>
<td>Indicator for association between MU $i$ and FBS $j$.</td>
</tr>
<tr>
<td>$a_j$</td>
<td>User association profile at FBS $j$.</td>
</tr>
<tr>
<td>$\mathcal{A}_j$</td>
<td>Feasible set of association profiles at base station $j$.</td>
</tr>
<tr>
<td>$\mathcal{A}'_j$</td>
<td>Set of association profiles with which FBS $j$ obtains more resources than the maximum amount it needs.</td>
</tr>
<tr>
<td>$B$</td>
<td>Bandwidth of each channel.</td>
</tr>
<tr>
<td>$c_i(\cdot)$</td>
<td>Function that characterizes the energy consumption of user $i$’s device.</td>
</tr>
<tr>
<td>$d$</td>
<td>Distance between MBS and the region center where MUs and FBSs are located.</td>
</tr>
<tr>
<td>$f_i(\cdot)$</td>
<td>Satisfaction function of user $i$ defined over $[0, +\infty)$.</td>
</tr>
<tr>
<td>$g_{ij}$</td>
<td>Average channel gain from user $i$ to base station $j$ due to path loss.</td>
</tr>
<tr>
<td>$h_{ij}$</td>
<td>Instantaneous small scale channel fading from user $i$ to base station $j$.</td>
</tr>
<tr>
<td>$I$</td>
<td>Set of base stations including MBS and FBSs.</td>
</tr>
<tr>
<td>$I_k$</td>
<td>Set of base stations to which channel $k$ has been allocated.</td>
</tr>
<tr>
<td>$J$</td>
<td>Total number of FBSs in the system.</td>
</tr>
<tr>
<td>$L$</td>
<td>Maximum number of users an FBS can serve.</td>
</tr>
<tr>
<td>$N$</td>
<td>Set of available channels for the system.</td>
</tr>
<tr>
<td>$N$</td>
<td>Total number of channels for the system.</td>
</tr>
<tr>
<td>$n_j$</td>
<td>Number of channels allocated to base station $j$.</td>
</tr>
<tr>
<td>$n_j^c$</td>
<td>Optimal number of channels allocated to base station $j$ under closed access.</td>
</tr>
<tr>
<td>$P_i$</td>
<td>User $i$’s transmission power.</td>
</tr>
<tr>
<td>$r_{ij}$</td>
<td>Indicator for access request from MU $i$ to FBS $j$.</td>
</tr>
<tr>
<td>$\overline{R}<em>{ij}, \tilde{R}</em>{ij}$</td>
<td>Average data rate and approximate data rate from user $i$ to base station $j$.</td>
</tr>
<tr>
<td>$\mathcal{S}^M$</td>
<td>Set of MUs in the system.</td>
</tr>
<tr>
<td>$\mathcal{S}_j^m, \mathcal{S}_j^f$</td>
<td>Set of MUs and set of FUs associated with base station $j$, respectively.</td>
</tr>
<tr>
<td>$T$</td>
<td>Number of time slots in one transmission period.</td>
</tr>
<tr>
<td>$u_{ij}$</td>
<td>Utility of user $i$ associated with base station $j$.</td>
</tr>
<tr>
<td>$\Delta u_{ij}$</td>
<td>Utility increment of user $i$ associated with FBS $j$ given one additional time slot.</td>
</tr>
<tr>
<td>$U_j$</td>
<td>Total utility at base station $j$.</td>
</tr>
<tr>
<td>$\Delta U_j$</td>
<td>Base station $j$’s utility increment given one additional channel.</td>
</tr>
<tr>
<td>$w_{ij}$</td>
<td>The proportion of time that user $i$ is allocated one channel from base station $j$.</td>
</tr>
<tr>
<td>$w_{i0}^{c}$</td>
<td>Optimal resource sharing factor for MU $i$ under closed access.</td>
</tr>
<tr>
<td>$x^k_j$</td>
<td>Channel allocation indicator of channel $k$ for base station $j$.</td>
</tr>
<tr>
<td>$\delta(k, x^k)$</td>
<td>Interference condition indicator for channel $k$ given channel allocation profile $x^k$.</td>
</tr>
<tr>
<td>$\Delta \varphi_j$</td>
<td>Additional channels that base station $j$ can obtain from MBS.</td>
</tr>
<tr>
<td>$\Phi_i(\cdot)$</td>
<td>Satisfaction function of user $i$ defined over $[R_i^{\min}, +\infty)$.</td>
</tr>
<tr>
<td>$\epsilon_j$</td>
<td>Access ratio for FBS $j$.</td>
</tr>
<tr>
<td>$\theta$</td>
<td>Interference threshold at FBSs.</td>
</tr>
<tr>
<td>$\beta$</td>
<td>Weighting factor in utility function.</td>
</tr>
<tr>
<td>$\sigma^2$</td>
<td>Additive white Gaussian noise power.</td>
</tr>
<tr>
<td>$\xi_j$</td>
<td>Maximum interference allowed at base station $j$.</td>
</tr>
</tbody>
</table>
Since the MUs are close to the FBSs, to avoid the cross-tier interference (from the MUs to the FBSs and from the FUs to the MBS), the scheduler allocates orthogonal channels to the macrocell tier and the femtocell tier. However, in the femtocell tier, the same channel may be reused by multiple femtocells under the condition that the maximum aggregate interference over that channel at each FBS is less than a predefined threshold $\theta$. Users associated with the same base station may share the available resources using time division multiple access, and at most one user in a femtocell can access a particular channel at a time slot.

Each FBS can serve at most $L$ users and is configured to operate in the hybrid access mode. In this access mode, the FBS may allow some MUs to access it but reserves a proportion of the available resources for its own FUs. To facilitate the access control process, we introduce an access ratio $0 \leq \epsilon_j < 1$, which denotes the maximum proportion of available resources that FBS $j \in \mathcal{I}\{0\}$ can allocate to the MUs associated with it. Note that $\epsilon_j = 0$ corresponds to closed access mode where the FBSs only serve their own FUs.

We assume users’ devices have the capability to sense the signal quality from nearby base stations and may send requests to access the FBS that provides the best signal quality. We consider each user requests a service (e.g., video chatting) that requires a minimum average data rate (over a transmission period) to guarantee its QoS. We assume the total amount of resources in the system is sufficient to satisfy the QoS requirements for all users. We consider that there is a network configuration phase to setup system parameters prior to users’ data transmission, which includes access control, resource allocation, and power management at the base stations and users. Such network configuration is performed periodically, i.e., every $T$ time slots.
3.3.1 Utility Function

In this subsection, we define utility functions to characterize the objectives at the users and base stations. In practice, a wireless user prefers receiving high QoS with low energy consumption. Therefore, we define the utility function for user $i \in S_j$ who is associated with base station $j \in \mathcal{I}$ as

$$u_{ij} = f_i(\overline{R}_{ij}) - c_i(\overline{P}_i),$$

(3.1)

where $f_i(\cdot)$ characterizes the satisfaction of user $i$ with respect to the average data rate over one transmission period ($\overline{R}_{ij}$), and $c_i(\cdot)$ is a function of the average transmission power ($\overline{P}_i$) that characterizes the energy consumption of user $i$’s device.

According to [83], the satisfaction of a user is an increasing function of the data rate, which has a decreasing marginal improvement as the data rate increases. Such property can be modeled using concave functions [84] [85]. To characterize users’ satisfaction under different QoS requirements, we consider a satisfaction function $f_i(\cdot)$ as shown in Fig. 3.2. For user $i$ associated with base station $j$, we have

$$f_i(\overline{R}_{ij}) = \begin{cases} \Phi_i(\overline{R}_{ij}), & \text{if } \overline{R}_{ij} \geq R_i^{\min}, \\ 0, & \text{otherwise,} \end{cases}$$

(3.2)

where $\Phi_i(\cdot)$ is a concave function defined in $[R_i^{\min}, \infty)$, and $R_i^{\min}$ is the minimum average data rate required to guarantee the basic QoS for user $i$. To simplify the analysis, we assume users have the same satisfaction at their QoS threshold, where $\Phi_i(R_i^{\min}) = C, \forall i \in S_j, \forall j \in \mathcal{I}$ and $C$ is a constant that guarantees positive utility for users. Equation (3.2) implies that user $i$ gains nothing when its data rate is less than the basic QoS requirement. Once the basic QoS is guaranteed, the user’s satisfaction becomes a concave function of the data rate. Equation (3.2) can be used to characterize the user’s satisfaction when running delay.

53
Figure 3.2: User’s satisfactory function $f_i(\bar{R}_{ij})$ with respect to the average data rate $\bar{R}_{ij}$. $R_{ij}^{\min}$ is the minimum rate requirement, $C$ is a constant, and $\Phi_i(\cdot)$ is a concave function defined in $[R_{ij}^{\min}, +\infty)$.

sensitive applications such as video chatting. Note that the user’s satisfaction function may vary with respect to different applications. The analysis and results in this chapter can be applied to the satisfaction function defined in (3.2) with any concave function $\Phi_i(\cdot)$.

We assume each user chooses a fixed transmission power from $[0, P_{\max}]$ to optimize their utilities during the transmission phase ($T$ time slots), that is $\bar{P}_i = P_i$ for $i \in \mathcal{S}_j, j \in \mathcal{I}$. We propose to use

$$c_i(P_i) = \beta w_{ij} P_i,$$

(3.3)

where $\beta$ is a constant weighting factor and $w_{ij} \in \{\frac{1}{T}, \frac{2}{T}, \ldots, 1\}$ is the resource sharing variable, which is defined as the proportion of time that user $i \in \mathcal{S}_j$ is granted access to base station $j \in \mathcal{I}$ during the transmission phase. $\beta$ is used to characterize the importance of the energy consumption in the utility function. It can be seen that when $\beta$ is small such that $c_i(P_i) < f_i(\bar{R}_{ij})$, users care more about data rate rather than their energy consumption. However, when $\beta$ is sufficiently large such that $c_i(P_i) > f_i(\bar{R}_{ij})$, energy
consumption becomes more critical when making the transmission decisions. Therefore, we can achieve different optimization objectives by adjusting the weighting factor $\beta$. Note that in the configuration phase, the actual data rate of a user for the transmission phase cannot be achieved since the channel gain in the future $T$ time slots is unknown. Therefore, we use an approximate data rate to calculate the utility when optimizing the decisions.

Based on the channel capacity in [68], we approximate the average data rate for user $i$ who communicates with base station $j$ using the average channel gain between them, where the approximate data rate is

\[
\tilde{R}_{ij} = w_{ij} B \log_2 \left( 1 + \frac{\mathbb{E}[g_{ij}|h_{ij}|^2 P_i]}{\sigma^2 + \xi_j} \right)
\]

\[
= w_{ij} B \log_2 \left( 1 + \frac{g_{ij} P_i}{\sigma^2 + \xi_j} \right),
\]

where $g_{ij}$ is the average channel gain from user $i$ to base station $j$, $h_{ij}$ is the corresponding small scale fading that follows a complex Gaussian distribution with zero mean and unit variance, $\sigma^2$ is the additive white Gaussian noise power, and $\xi_j$ is the maximum interference allowed at base station $j$. Here, we use the fact that $\mathbb{E}[|h_{ij}|^2] = 1$ for complex Gaussian variable $h_{ij}$ with unit variance. Note that in (3.4) we use the maximum allowed interference instead of the actual interference, since the actual interference experienced at a base station is unknown prior to the transmission phase. Since orthogonal channels are allocated to the macrocell tier and femtocell tier, there is no interference at the MBS. Therefore, we have

\[
\xi_j = \begin{cases} 
\theta, & \text{if } j \in \mathcal{I} \setminus \{0\}, \\
0, & \text{otherwise.}
\end{cases}
\]

Thus, the approximate data rate of user $i$ to base station $j$ is a function of the resource sharing variable $w_{ij}$ and the transmission power $P_i$. In the remainder of this chapter, we
use the approximate data rate to calculate the utility for each user when designing the network configuration mechanism.

We assume the MBS aims to optimize the total utility of the MUs. Each FBS allocates the resource to maximize the total utility of its own FUs, while guaranteeing the QoS of all its associated users (FUs and MUs). Since users receive wireless service from the same provider, and the wireless spectrum is limited, the operator may make channel allocation decisions to optimize the performance of the whole network, i.e., to maximize the system throughput [78] or total utility [86]. We consider the operator’s objective at the scheduler is to optimize the total utility of all users in the system. Note that the proposed configuration mechanism and corresponding analysis in this chapter can also be applied to other utility functions with proper adjustment.

### 3.3.2 Network Configuration Mechanism

We aim to design a network configuration mechanism that includes access control, resource allocation, and power management between base stations and users. We consider the FBSs perform access control at the beginning and the scheduler allocates the channels to the FBSs adaptively based on their access control decisions. This is to encourage the FBSs to use hybrid access mode to serve MUs who receive low QoS from the MBS. Specifically, if an FBS grants access to some MUs, it may obtain more channels and thus may improve its own utility. Note that to determine the transmission power that optimizes the utility, a user may require the knowledge of which base station it is associated with and the amount of resource it is granted from that base station. Therefore, it is reasonable to consider power management after the resource allocation process.

Based on the above discussion, we propose a network configuration mechanism that follows the procedure as shown in Fig. 3.3, which consists of five stages. The first two
stages correspond to the access control process at the FBSs. In Stage I, after receiving the access requests from the MUs, each FBS $j$ ($j \in \mathcal{I}\{0\}$) determines its association profile $\mathbf{a}_j = (a_{ij}, i \in \mathcal{S}^M)$, where $a_{ij} = 1$ indicates that MU $i$ is granted access to FBS $j$. Note that each MU can only be associated with one base station, which leads to $\sum_{j\in\mathcal{I}\{0\}} a_{ij} \leq 1$. Then, FBS $j$ determines its access ratio $\epsilon_j$ in Stage II. After the access control process, the FBSs send the access control decisions (including the association decision profile $\mathbf{A} = \{\mathbf{a}_j, j \in \mathcal{I}\{0\}\}$ and the access ratios $\mathbf{e} = (\epsilon_j, \forall j \in \mathcal{I}\{0\})$) to the scheduler. In Stage III, the scheduler performs channel allocation. We define $x_j^k \in \{0, 1\}$ as the indicator whether channel $k$ is allocated to base station $j$ ($x_j^k = 1$) or not ($x_j^k = 0$). The scheduler determines the set of channel allocation profiles $\mathbf{x} = (\mathbf{x}_j, \forall j \in \mathcal{I})$, where $\mathbf{x}_j = (x_j^k, \forall k \in \mathcal{N})$ denotes the channel allocation profile at base station $j$. Then, the MBS and the FBSs allocate their available resources (channels and time slots) to their associated users in Stage IV, i.e., base station $j \in \mathcal{I}$ determines its resource allocation profile $\mathbf{w}_j = (w_{ij}, \forall i \in \mathcal{S}_j)$. Finally, each user $i$ determines its transmission power $P_i$ based on the available resource.
from its associated base station.

The proposed network configuration mechanism can be viewed as a multi-stage decision making process, where base stations and users make decisions sequentially. Our objective is to design efficient decision making strategies for the base stations and users in order to optimize their own utilities.

### 3.3.3 Multi-level Optimization Problem

In the proposed network configuration mechanism, base stations and users have different objectives. Therefore, instead of formulating a single utility maximization problem which only targets one objective, in this chapter, we adopt a multi-level optimization approach [87] and formulate the network configuration process as a five-level optimization problem. The multi-level optimization approach models each decision making stage as an optimization problem with a specific objective function. It finds the optimal decision at one stage by considering the decisions made in the previous stages and the decisions to be made in the following stages. We define \( r_{ij} \) as the indicator whether FBS \( j \) receives a request from MU \( i \), where \( r_{ij} = 1 \) indicates the request is received and \( r_{ij} = 0 \) otherwise.

We define the set of feasible association decisions for FBS \( j \in \mathcal{I}\{0\} \) as

\[
\mathcal{A}_j \triangleq \{ a_j | a_{ij} \leq r_{ij}, \forall i \in \mathcal{S}^M, \sum_{i \in \mathcal{S}^M} a_{ij} \leq L - |\mathcal{S}_j^f| \},
\]

where \(| \cdot |\) is the cardinality of the set. We further denote \( P_j = (P_i, i \in \mathcal{S}_j) \) for \( j \in \mathcal{I}\{0\} \), and define \( \epsilon_0 = 0 \) and \( a_0 = (a_{i0}, i \in \mathcal{S}_0) \). We define \( U_j \) as the utility of base station \( j \in \mathcal{I} \).

As mentioned in Section 3.3.1, the utility of MBS is the total utility of MUs, which is \( U_0 = \sum_{i \in \mathcal{S}^M} u_{i0} \). The utility of FBS \( j \in \mathcal{I} \{0\} \) is the total utility of its FUs, which is \( U_j = \sum_{i \in \mathcal{S}_j^f} u_{ij} \). Note that a user’s utility depends on its transmission power \( P_i \) and the amount of allocated resources \( w_{ij} \), while the determination of \( P_i \) and \( w_{ij} \) is also related.
Chapter 3. Network Configuration for Two-tier Heterogeneous Networks

to the access control decision \(a_j, \epsilon_j\) and the channel allocation decision \(x_j\). Therefore, the utility at each base station \(j \in \mathcal{I}\) is a function of the variables \(a_j, \epsilon_j, x_j, w_j\) and \(P_j\). In the following, we use \(U_j(a_j, \epsilon_j, x_j, w_j, P_j)\) as the utility function at base station \(j\) during the problem formulation. Then, the multi-level optimization problem, denoted as problem \(\mathcal{P}\), can be represented as

Level I: For \(j \in \mathcal{I}\backslash\{0\}\),

\[
\max_{a_j} \ U_j(a_j, \epsilon_j, x_j, w_j, P_j)
\]

subject to \(a_j \in \mathcal{A}_j\),

where \(\epsilon_j \ (j \in \mathcal{I}\backslash\{0\})\) is the solution of

Level II: For \(j \in \mathcal{I}\backslash\{0\}\),

\[
\max_{\epsilon_j} \ U_j(a_j^*, \epsilon_j, x_j, w_j, P_j)
\]

subject to \(\epsilon_j^{\min} \leq \epsilon_j < 1\),

where \(x_j \ (j \in \mathcal{I}\backslash\{0\})\) is the solution of

Level III:

\[
\max_{x_j, j \in \mathcal{I}} \sum_{j \in \mathcal{I}} U_j(a_j^*, \epsilon_j^*, x_j, w_j, P_j)
\]

subject to \(n_j^{\min} \leq \sum_{k \in \mathcal{N}} x_j^k \leq n_j^{\max}, \ \forall \ j \in \mathcal{I}\),

\[
x_j^k \sum_{l \in \mathcal{J}\backslash\{j, 0\}} x_l^k \max_{i \in \mathcal{S}_l} \{g_{ij} P_{\text{max}}\} \leq \theta, \ \forall \ k \in \mathcal{N}, j \in \mathcal{I}\backslash\{0\},
\]

where \(w_j \ (j \in \mathcal{I})\) is the solution of

Level IV: For \(j \in \mathcal{I}\),

\[
\max_{w_j} \ U_j(a_j^*, \epsilon_j^*, x_j^*, w_j, P_j)
\]

subject to \(\sum_{i \in \mathcal{S}_j} w_{ij} \leq \sum_{k \in \mathcal{N}} x_j^k\),

\[j(\sum_{i \in \mathcal{S}_j} w_{ij} - \epsilon_j \sum_{k \in \mathcal{N}} x_j^k) \leq 0,\]

\[w_{ij} \geq w_{ij}^{\min}, \ \forall \ i \in \mathcal{S}_j,\]
where $\mathbf{P}_j$ ($j \in \mathcal{I}$) is the solution of

Level V: For $i \in \mathcal{S}_j$, $j \in \mathcal{I}$,

$$\maximize_{P_i} u_{ij}(w_{ij}^*, P_i)$$

subject to $P_i \in [0, P_{\text{max}}]$,

where $a_j^*, \epsilon_j^*, x_j^*$ and $w_{ij}^*$ represent the optimal decisions at the corresponding optimization level. In the above formulation, at each level, we need to solve an optimization problem taking into account the parameter(s) obtained from the upper level(s) and the optimality of problem(s) in the following level(s). For example, in Level II, FBS $j$ determines its access ratio $\epsilon_j$ based on the user association decision $a_j$ obtained in Level I and the other optimal decisions (i.e., $x_j$) obtained from lower levels. The determination of access ratio in this stage is important since $\epsilon_j$ works as an indicator for how many channels should be allocated to FBS $j$ in order to guarantee the minimum QoS requirement(s) for the MU(s) accepted by FBS $j$. We require $\epsilon_j$ be no smaller than a certain value, denoted as $\epsilon_{\text{min}}^j$. It also indicates how much resources the FBS would like to share with its associated MUs. In Level III, $n_{\text{min}}^j$ and $n_{\text{max}}^j$ denote the minimum and maximum number of channels that can be allocated to FBS $j$, respectively. Given the association decisions from the FBSs, $n_{\text{min}}^j$ and $n_{\text{max}}^j$ should be properly selected to guarantee the feasibility of optimization in the following levels. The channel allocation decision at this level also satisfies the maximum allowed interference constraint at each FBS. In Level IV, $w_{ij}^\text{min}$ represents the minimum resource that user $i$ requires from base station $j$ to guarantee its QoS requirement. The determination of the parameters ($\epsilon_j^\text{min}, n_{\text{min}}^j, n_{\text{max}}^j$, and $w_{ij}^\text{min}$) is discussed in Section 3.4. The solution for each optimization problem corresponds to the optimal decision(s) for the decision maker(s) at that stage. Thus, our objective becomes designing efficient algorithms to find the optimal solution for each optimization level.
3.4 Analysis of the Multi-level Optimization Problem

In this section, we analyze the multi-level optimization problem and design efficient algorithms to achieve the optimal solution at each level. The multi-level optimization problem can be analyzed by a bottom-up approach [87], which follows the order from the lowest to the highest level sequentially.

3.4.1 Power Management at Users

We first consider the last level of problem $\mathbb{P}$. Our objective is to find the optimal transmission power for each user $i \in S_j$ given the resource sharing variable $w_{ij}$ from its associated base station $j \in \mathcal{I}$. In this level, the utility of a user becomes a function of the transmission power. The corresponding optimization problem is

$$
\begin{align*}
\text{maximize} & \quad u_{ij}(w_{ij}, P_i) \\
\text{subject to} & \quad P_i \in [0, P_{\text{max}}].
\end{align*}
$$

(3.7)

We analyze problem (3.7) with respect to different values of $w_{ij}$. Note that the minimum amount of resources that user $i$ requires from base station $j$ to guarantee its QoS can be calculated as

$$
w_{ij}^{\text{min}} = \frac{P_i^{\text{min}}}{B \log_2 \left( 1 + \frac{g_{ij} P_{\text{max}}}{\sigma^2 + \xi_j} \right)}.
$$

(3.8)

First, when $w_{ij} < w_{ij}^{\text{min}}$, according to equations (3.1), (3.2) and (3.3), we have $u_{ij}(P_i) = -\beta w_{ij} P_i$, and the optimal power is $P_i^* = 0$ in this scenario. Next, when $w_{ij} \geq w_{ij}^{\text{min}}$, we define $P_i^{\text{min}} \leq P_{\text{max}}$ as user $i$’s minimum power that satisfies its QoS requirement, which is given by $P_i^{\text{min}} = \frac{\sigma^2 + \xi_j}{g_{ij}} (2^{P_i^{\text{min}}/w_{ij} B} - 1)$. In this scenario, to solve problem (3.7),...
we only need to consider the feasible region \([P_i^{\text{min}}, P_{\text{max}}]\). For \(P_i \in [P_i^{\text{min}}, P_{\text{max}}]\), we have

\[ u_{ij}(P_i) = \Phi_i(\tilde{R}_{ij}) - \beta w_{ij} P_i \]

and

\[
\frac{du_{ij}}{dP_i} = \frac{d\Phi_i}{d\tilde{R}_{ij}} \frac{d\tilde{R}_{ij}}{dP_i} - \beta w_{ij} \tag{3.9}
\]

Note that \(\Phi_i(\cdot)\) is a concave increasing function of \(\tilde{R}_{ij}\), which implies that \(\frac{d\Phi_i}{d\tilde{R}_{ij}}\) is a decreasing function of \(\tilde{R}_{ij}\) (or \(\frac{d^2 \Phi_i}{d\tilde{R}_{ij}^2} < 0\)). Since \(\tilde{R}_{ij}\) is increasing with \(P_i\), \(\frac{d\Phi_i}{d\tilde{R}_{ij}}\) is a decreasing function of \(P_i\). From (3.9), we conclude that \(\frac{du_{ij}}{dP_i}\) is also decreasing with \(P_i\) in \((P_i^{\text{min}}, P_{\text{max}})\). Based on the decreasing property of \(\frac{du_{ij}}{dP_i}\), if \(\frac{du_{ij}}{dP_i} \bigg|_{P_i = P_i^{\text{min}}} \leq 0\), we have \(\frac{du_{ij}}{dP_i} < 0\) for \(P_i \in (P_i^{\text{min}}, P_{\text{max}})\). Therefore, the optimal power in this case is \(P_i^* = P_i^{\text{min}}\).

On the other hand, if \(\frac{du_{ij}}{dP_i} \bigg|_{P_i = P_i^{\text{min}}} > 0\), since \(\frac{du_{ij}}{dP_i} \bigg|_{P_i \to \infty} = -\beta w_{ij} < 0\) and \(\frac{du_{ij}}{dP_i}\) is decreasing with \(P_i\), equation \(\frac{du_{ij}}{dP_i} = 0\) has a unique root \(P_i = \tilde{P}_i\) in \((P_i^{\text{min}}, \infty)\), where we have \(\frac{du_{ij}}{dP_i} > 0\) in \((P_i^{\text{min}}, \tilde{P}_i)\) and \(\frac{du_{ij}}{dP_i} < 0\) in \((\tilde{P}_i, \infty)\). Therefore, in this case, the optimal power is \(P_i^* = \min(\tilde{P}_i, P_{\text{max}})\). In summary, the optimal solution to problem (3.7) can be represented as

\[
P_i^* = \begin{cases} 
I\{\frac{du_{ij}}{dP_i} \bigg|_{P_i = P_i^{\text{min}}} \leq 0\} P_i^{\text{min}} + I\{\frac{du_{ij}}{dP_i} \bigg|_{P_i = P_i^{\text{min}}} > 0\} \min(\tilde{P}_i, P_{\text{max}}), & \text{if } w_{ij} \geq w_{ij}^{\text{min}}, \\
0, & \text{otherwise},
\end{cases} \tag{3.10}
\]

where indicator function \(I_X = 1\) if \(X\) is true, and is equal to zero otherwise. It can be seen that given the resource sharing factor \(w_{ij}\) from base station \(j \in \mathcal{I}\), the optimal power management decision for user \(i \in \mathcal{S}_j\) is provided in (3.10).
3.4.2 Resource Allocation at Base Stations

Level IV of problem $\mathbb{P}$ corresponds to resource allocation at the base stations. As mentioned in Section 3.3, the channel gain between a user and a base station is i.i.d.. Therefore, when optimizing the resource allocation decisions, we only determine the resource sharing variables $w_{ij}$, but do not consider which channel or time slots are allocated to each user. After determining the resource sharing variables, the number of time slots for each user is fixed. The set of time slots for each user can be determined by applying a scheduling algorithm, such as round robin scheduling algorithm or random scheduling algorithm. We first consider resource allocation at FBS $j$ given its access control decision ($a_j$ and $\epsilon_j$) and the channel allocation decision $x_j$ from the scheduler. Note that we require the FBS’s decision guarantee the QoS requirements of all its associated users, and we assume such decision is feasible. The optimal resource allocation decision $w_j = (w_{ij}, \forall i \in S_j)$ that maximizes FBS $j$’s utility can be found by solving the following problem:

$$\begin{align*}
\text{maximize} \quad & \sum_{i \in S_j^f} w_{ij}(w_{ij}, P_i^*(w_{ij})) \\
\text{subject to} \quad & \sum_{i \in S_j^m} w_{ij} \leq \epsilon_j \sum_{k \in N} x_{kj}^k, \\
& \sum_{i \in S_j^f} w_{ij} + \sum_{i \in S_j^m} w_{ij} \leq \sum_{k \in N} x_{kj}^k, \\
& w_{ij} \geq w_{ij}^{\min}, \quad \forall i \in S_j, \\
& w_{ij} \in \{1, \frac{2}{T}, \ldots, 1\}, \quad \forall i \in S_j,
\end{align*}$$

(3.11)

where $P_i^*(w_{ij})$ is the optimal transmission power for user $i$ given $w_{ij}$. The first constraint indicates that the FBS allocates at most $\epsilon_j$ of the total resource to the MUs. The second constraint implies that the total amount of resources allocated to users should not exceed the total number of channels available at the FBS. The third constraint guarantees the QoS of all users associated with this FBS. The last constraint implies that a user can only obtain
an integer number of time slots and cannot access more than one channel simultaneously. Since $P_i^*(w_{ij})$ depends on $w_{ij}$, from (3.9) and (3.10), it can be seen that problem (3.11) is a nonlinear discrete optimization problem, which is hard to solve in general. In the rest of this section, we design an efficient algorithm to achieve the optimal solution.

Since the objective function in problem (3.11) is only related to FUs associated with FBS $j$, problem (3.11) can be solved in two steps, where we first determine the resource sharing variables for MUs associated with FBS $j$, and then we determine the resource sharing variables for FUs. Intuitively, to maximize FUs’ total utility, the FBS should allocate as much resources as possible to its FUs, or equivalently, to minimize the amount of resources allocated to the MUs while satisfying their QoS constraints. Therefore, in the first step, the optimal resource sharing variables for the MUs can be found by solving the following subproblem:

$$\min_{w_{ij}, i \in S^m_j} \sum_{i \in S^m_j} w_{ij}$$
subject to
$$\sum_{i \in S^m_j} w_{ij} \leq \epsilon_j \sum_{k \in \mathcal{N}} x^k_j,$$
$$w_{ij} \geq w_{ij}^{\min}, \quad \forall \ i \in S^m_j,$$
$$w_{ij} \in \{\frac{1}{T}, \frac{2}{T}, \ldots, 1\}, \quad \forall \ i \in S^m_j.$$

(3.12)

In the second step, with the solution to subproblem (3.12), $w_{ij}^*, \forall \ i \in S^m_j$, problem (3.11) is transformed into the following subproblem:

$$\max_{w_{ij}, i \in S^f_j} \sum_{i \in S^f_j} u_{ij}(w_{ij}, P_i^*(w_{ij}))$$
subject to
$$\sum_{i \in S^f_j} w_{ij} \leq \sum_{k \in \mathcal{N}} x^k_j - \sum_{i \in S^m_j} w_{ij}^*, $$
$$w_{ij} \geq w_{ij}^{\min}, \quad \forall \ i \in S^f_j,$$
$$w_{ij} \in \{\frac{1}{T}, \frac{2}{T}, \ldots, 1\}, \quad \forall \ i \in S^f_j.$$

(3.13)
Algorithm 3.1 Optimal resource allocation algorithm at FBS $j$ with fixed user association

1: **Basic resource allocation:**
2: Calculate the minimum amount of resource required by each user, $w_{ij}^\text{min}$, $\forall \ i \in S_j$.
3: Set $w_{ij} := w_{ij}^\text{min}$, $\forall \ i \in S_j$.

4: **Remaining resource allocation:**
5: Determine the total remaining time slots $T_R := T(\sum_{k \in N} x_j^k - \sum_{i \in S_j} w_{ij}^\text{min})$.
6: Calculate the initial value of $\Delta u_{ij}(w_{ij} + 1/T)$, $\forall \ i \in S_f^j$.
7: **Repeat**
8: Find the set of FUs, $S_t$, that satisfies $w_{ij} < 1, \Delta u_{ij}(w_{ij} + 1/T) > 0, \forall \ i \in S_t$.
9: if $S_t \neq \emptyset$
10: Find user $i \in S_t$ that has the largest $\Delta u_{ij}(w_{ij} + 1/T)$ and set $w_{ij} := w_{ij} + 1/T, T_R := T_R - 1$.
11: Update $\Delta u_{ij}(w_{ij} + 1/T)$ for user $i$.
12: endif
13: **Until** $T_R = 0$ or $S_t = \emptyset$.

By solving subproblem (3.13), we obtain the optimal resource sharing variables for the FUs. It can be seen that the solutions to subproblems (3.12) and (3.13) constitutes the solution to problem (3.11). We define $\Delta u_{ij}(w_{ij}) \overset{\Delta}{=} u_{ij}(w_{ij}, P^*_i(w_{ij})) - u_{ij}(w_{ij} - \frac{1}{T}, P^*_i(w_{ij} - \frac{1}{T}))$ as the changes of utility when one time slot is assigned to user $i$. Based on the previous discussion, we propose a two-step greedy allocation algorithm to find the optimal solution for (3.11), as shown in Algorithm 3.1. In the first step, we allocate the minimum number of time slots ($w_{ij}^\text{min} T$) for each user to satisfy their QoS constraints. Intuitively, this solves subproblem (3.12). In the second step, we allocate the remaining time slots one by one, where each time slot is allocated to the FU who has the largest utility increment ($\Delta u_{ij}(w_{ij})$). With Algorithm 3.1, the following result can be obtained.

**Theorem 3.1** The resource allocation profile $w_j$ obtained using Algorithm 3.1 constitutes the optimal solution to problem (3.11).

The proof of Theorem 3.1 is provided in Appendix B. For the MBS, the resource allocation algorithm is similar to that at the FBS, except that the MBS only serves MUs. This algorithm can be obtained by replacing the terms FU by MU, and $S_f^j$ by $S_0$, respectively, in Algorithm 3.1. Thus, we have found the desired strategy for FBSs and
MBS in Level IV. Note that we can apply other objective functions to realize alternative optimization objectives. For example, we can use \( \sum_{i \in S_j} \log(R_{ij}) \) to achieve proportional fairness of data rate among the users. The optimality of the proposed algorithm is still valid as long as the objective function is concave with respect to the data rate.

### 3.4.3 Channel Allocation at the Scheduler

Next, we consider the channel allocation problem in Level III of problem \( \mathbb{P} \). In this level, the scheduler allocates the available channels \( (N) \) to each base station in order to maximize the total system utility. The channel allocation decision should guarantee the feasibility of the decisions in the following levels, which is to guarantee the QoS requirement for each user in the system. To achieve this, we require that the number of channels allocated to FBS \( j \in I \), defined as \( n_j \triangleq \sum_{k \in N^c} x_{kj}^k \), satisfies \( n_j^{\min} \leq n_j \leq n_j^{\max} \). We define \( n_j^{\min} \) and \( n_j^{\max} \) based on the access control decisions at base station \( j \) as follows. First, since we assume each user can only access one channel at a time, the total number of channels allocated to a base station should not exceed the number of its associated users. Therefore, it is reasonable to define \( n_j^{\max} = |S_j| \). For base station \( j \) which only serves its own subscribed users (either MUs or FUs), \( n_j^{\min} \) is defined as \( n_j^{\min} = \lceil \sum_{i \in S_j} w_{ij}^{\min} \rceil \), where \( \lceil \cdot \rceil \) is the ceiling function. For FBS \( j \), which serves both FUs and MUs, to guarantee the QoS of all its associated users, \( n_j^{\min} \) should satisfy

\[
\epsilon_j n_j^{\min} \geq \sum_{i \in S_j^m} w_{ij}^{\min}, \tag{3.14}
\]

and

\[
(1 - \epsilon_j) n_j^{\min} \geq \sum_{i \in S_j^f} w_{ij}^{\min}. \tag{3.15}
\]
That is

\[
\begin{align*}
n_{j}^{\text{min}} & \geq \max \left( \left\lceil \frac{1}{\epsilon_j} \sum_{i \in S_{j}^{m}} w_{ij}^{\text{min}} \right\rceil, \left\lceil \frac{1}{1-\epsilon_j} \sum_{i \in S_{j}^{l}} w_{ij}^{\text{min}} \right\rceil \right). 
\end{align*}
\]

(3.16)

Note that if FBS \( j \) intentionally selects a small \( \epsilon_j \) that is close to zero, \( n_{j}^{\text{min}} \) may become a large number, which makes the allocation problem infeasible. To prevent such behavior, we set an upper bound for \( n_{j}^{\text{min}} \) as follows. We denote \( n_{j}^{c} \) as the optimal number of channels allocated to base station \( j \) when all base stations use closed access mode, and represent the corresponding optimal resource sharing variable for MU \( i \in S^{M} \) (associated with MBS) as \( w_{i0}^{c} \). Then, for FBS \( j \) who accepts to serve the set of MUs \( S_{j}^{m} \), the minimum number of channels to be allocated satisfies

\[
\begin{align*}
n_{j}^{\text{min}} & \leq n_{j}^{c} + \left\lfloor \sum_{i \in S_{j}^{m}} w_{i0}^{c} \right\rfloor. 
\end{align*}
\]

(3.17)

Inequality (3.17) implies that if FBS \( j \) accepts to serve MUs, it is guaranteed to obtain up to \( \left\lfloor \sum_{i \in S_{j}^{m}} w_{i0}^{c} \right\rfloor \) additional channels from the scheduler (depending on the selected access ratio). Note that allocating these additional channels to the FBS does not reduce the number of channels allocated to other base stations when the set of MUs switch from accessing the MBS to accessing the FBS.

Based on the previous discussion, we determine \( n_{j}^{\text{min}} \) as

\[
\begin{align*}
n_{j}^{\text{min}} = \begin{cases} 
\min\{ \max \left( \left\lceil \frac{1}{\epsilon_j} \sum_{i \in S_{j}^{m}} w_{ij}^{\text{min}} \right\rceil, \left\lceil \frac{1}{1-\epsilon_j} \sum_{i \in S_{j}^{l}} w_{ij}^{\text{min}} \right\rceil \right), 
n_{j}^{c} + \left\lfloor \sum_{i \in S_{j}^{m}} w_{i0}^{c} \right\rfloor, 
\left\lceil \sum_{i \in S_{j}^{l}} w_{ij}^{\text{min}} \right\rceil, 
\end{cases} 
\end{align*}
\]

(3.18)

With the above definition, the optimal channel allocation strategy that maximizes the
system utility can be found by solving the following problem

$$\max_{x_{j}, j \in \mathcal{I}} \sum_{j \in \mathcal{I}} \sum_{i \in \mathcal{S}} u_{ij}(w_{ij}^{*}(x_{j}), P_{i}^{*}(w_{ij}^{*}(x_{j})))$$

subject to

$$n_{j}^{\min} \leq \sum_{k \in \mathcal{N}} x_{j}^{k} \leq n_{j}^{\max}, \forall j \in \mathcal{I},$$

$$x_{j}^{k} \sum_{l \in \mathcal{I}\backslash\{j, 0\}} x_{l}^{k} \max_{i \in \mathcal{S}_{l}} \{g_{ij} P_{\max}\} \leq \theta, \forall k \in \mathcal{N}, j \in \mathcal{I}\backslash\{0\},$$

$$x_{j}^{k} \in \{0, 1\}, \quad \forall k \in \mathcal{N}, j \in \mathcal{I},$$

(3.19)

where $P_{i}^{*}(w_{ij}^{*}(x_{j}))$ and $w_{ij}^{*}(x_{j})$ can be found for a given $x_{j}$ using algorithms in Sections 3.4.1 and 3.4.2, respectively. Note that $x_{j}^{k}$ takes integer values and the objective function in (3.19) does not have an explicit form with respect to $x_{j}^{k}$. Therefore, it is difficult to find the optimal solution. Similar to Algorithm 3.1, we propose a greedy allocation algorithm to find a suboptimal solution.

We define $\Delta U_{j}(n_{j}) = U_{j}(n_{j}) - U_{j}(n_{j} - 1)$ as the utility increment by assigning one additional channel to base station $j$. We define $\delta(k, \mathbf{x}^{k}) \in \{0, 1\}$ as the indicator whether the interference constraint for all FBSs over channel $k$ is satisfied or not, given the channel allocation profile $\mathbf{x}^{k} = (x_{j}^{k}, j \in \mathcal{I}\backslash\{0\})$. We further denote $\mathbf{x}^{k}|_{j}$ as the channel allocation profile obtained by setting $x_{j}^{k} = 1$ from $\mathbf{x}^{k}$, and $\delta(k, \mathbf{x}^{k}|_{j})$ indicates whether the interference constraints are satisfied when allocating channel $k$ to FBS $j$. We determine the value of $\delta(k, \mathbf{x}^{k}|_{j})$ as follows. At the beginning of the channel allocation process, we determine a $J \times J$ matrix $\Gamma^{f}$, where $\Gamma_{i,j}^{f}$ represents the maximum interference from users associated with FBS $i$ to FBS $j$. That is, $\Gamma_{i,j}^{f} = \max_{u \in \mathcal{S}_{i}} \{P_{u} g_{u,j}\}$. Note that $\Gamma^{f}$ characterizes the interference between any pair of femtocells and $\Gamma_{i,j}^{f}$ is different from $\Gamma_{j,i}^{f}$. We also maintain a $J \times N$ interference matrix $\Gamma^{c}$, where $\Gamma_{j,k}^{c}$ represents the aggregate interference experienced at FBS $j$ on channel $k$. To determine $\delta(k, \mathbf{x}^{k}|_{j})$ when allocating a channel $k$ to an FBS $j$, we first find the set of FBSs that have already been allocated channel $k$, denoted as $\mathcal{I}^{k}$.
Then, we calculate the temporary aggregate interference for FBS $i \in \mathcal{I}^k$ over channel $k$ assuming channel $k$ is allocated to FBS $j$, which is $\Gamma'_{i,k} = \Gamma^c_{i,k} + \Gamma^f_{j,i}$, where we add the new interference generated from users associated with FBS $j$ ($\Gamma^f_{j,i}$) to existing interference at FBS $i \in \mathcal{I}^k$ ($\Gamma^c_{i,k}$). We also calculate the maximum aggregate interference experienced at FBS $j$ from other femtocells on channel $k$, which is $\Gamma'_{j,k} = \sum_{i \in \mathcal{I}^k} \Gamma^f_{i,j}$. Then, the interference condition indicator can be determined as

$$
\delta(k, x^k|j) = \begin{cases} 
1, & \text{if } \Gamma'_{i,k} < \theta, \forall i \in \mathcal{I}^k \cup \{j\}, \\
0, & \text{otherwise.}
\end{cases} 
$$

(3.20)

Once channel $k$ is allocated to FBS $j$, we update $\Gamma^c_{i,k}$ and $\Gamma^f_{i,j}$ $\forall i, j \in \mathcal{I}$ accordingly. By introducing the indicator function $\delta(k, x^k|j)$, the proposed greedy channel allocation algorithm is shown in Algorithm 3.2. The main idea of Algorithm 3.2 is to allocate the channels one by one, where each channel is allocated to as many base stations as possible under the interference constraints. The available channels are indexed from 1 to $N$. Since MBS and FBSs use different set of channels, we allocate the channels to MBS starting from index $N$ in a decreasing order, while allocating channels to FBSs starting from index 1 in an increasing order.

Algorithm 3.2 contains two main steps. In the first step, we allocate the minimum number of channels to each base station so that $n_j = n^\text{min}_j$, $\forall j \in \mathcal{I}$. Specifically, we allocate $n^\text{min}_0$ channels, which indexed from $N$ to $N - n^\text{min}_0 + 1$, to the MBS in Line 3. Then, we allocate the minimum number of channels to FBSs in Lines 4 to 14 starting from channel indexed at 1, where for each channel, we try to allocate it to as many FBSs as possible by checking the interference constraint for each FBS (from Lines 8 to 12). After allocating the minimum number of channels, in the second step, we allocate the remaining channels one by one to the MBS and FBSs, respectively, from Lines 16 to 33. We introduce
Algorithm 3.2 Greedy channel allocation algorithm

1: Basic channel allocation:
2: Calculate $n^\text{min}_j$ and $n^\text{max}_j$, $\forall j \in \mathcal{I}$.
3: Set $x^k_0 := 1$, $\forall k = N - n^\text{min}_0 + 1, \ldots, N$.
4: Set $k := 1$.
5: Find the set of FBSs $\mathcal{I}_b$ that satisfies $n_j < n^\text{min}_j$, $\forall j \in \mathcal{I}_b$.
6: while $(\mathcal{I}_b \neq \emptyset)$
7:   Set $k := k + 1$.
8:   for each FBS $j \in \mathcal{I}_b$
9:     if $\delta(k, x^k_j) = 1$
10:        Set $x^k_j := 1$, $n_j := n_j + 1$.
11:   endif
12: endfor
13: Find the set of FBSs $\mathcal{I}_b$ that satisfies $n_j < n^\text{min}_j$, $\forall j \in \mathcal{I}_b$.
14: endwhile
15: Remaining channel allocation:
16: Set the starting point of remaining channels for FBS and MBS $n^\text{FBS} := 2$, $n^\text{MBS} := N - n^\text{min}_0$.
17: Calculate the initial value of $\Delta U_j(n_j + 1)$, $\forall j \in \mathcal{I}$ based on Algorithm 3.1.
18: Find the base station set $\mathcal{I}'_b$ that satisfies $n_j < n^\text{max}_j$ and $\Delta U_j(n_j + 1) > 0$.
19: while $(\mathcal{I}'_b \neq \emptyset)$ and $n^\text{FBS} < n^\text{MBS}$
20:   Find base station $j \in \mathcal{I}'_b$ that has the largest $\Delta U_j(n_j + 1)$.
21:   if $j = 0$ and $n^\text{MBS} > k$
22:      Set $x^0_0 := 1$, $n_0 := n_0 + 1$, $n^\text{MBS} := n^\text{MBS} - 1$.
23:      Update $\Delta U_0(n_0 + 1)$ by computing $U_0(n_0 + 1)$ and $U_0(n_0)$ according to Algorithm 3.1.
24:   else
25:      for FBS $j$ in decreasing order with respect to $\Delta U_j(n_j + 1)$
26:         if $\delta(n^\text{FBS}_j, x^{n^\text{FBS}_j}_j) = 1$ and $x^{n^\text{FBS}_j}_j = 0$
27:            Set $x^{n^\text{FBS}_j}_j := 1$, $n_j := n_j + 1$.
28:            Update $\Delta U_j(n_j + 1)$ by computing $U_j(n_j + 1)$ and $U_j(n_j)$ according to Algorithm 3.1.
29:         endif
30:      endfor
31:   endif
32: endwhile
33: Find the base station set $\mathcal{I}'_b$ that satisfies $n_j < n^\text{max}_j$ and $\Delta U_j(n_j + 1) > 0$, $\forall j \in \mathcal{I}'_b$.
34: endwhile
two variables $n_{\text{FBS}}$ and $n_{\text{MBS}}$ to denote the index of channels to be allocated to FBSs and MBS, respectively. We initialize $n_{\text{FBS}} = 2$ in Line 16. This is because in the basic allocation process we only consider those FBSs whose minimum requirements are not satisfied, and the channels indexed from 2 to $k$ may still be allocated to some FBSs without violating the interference constraints. We initialize $n_{\text{MBS}} = N - n_{0}^{\text{min}}$ and allocate channels to MBS following a decreasing order of channel index. Then, in each iteration, we find the base station with the largest utility increment assuming one additional channel is allocated to it in Line 20. If it is the MBS, we allocate channel $n_{\text{MBS}}$ to MBS in Lines 22 to 23. Otherwise, we allocate channel $n_{\text{FBS}}$ to as many FBSs as possible under the interference constraints from Lines 25 to 30 according to the order of utility increment at the FBSs. The allocation terminates when no more channels are available ($n_{\text{FBS}} \geq n_{\text{MBS}}$) or all the base stations have been fully allocated ($n_{j} = n_{j}^{\text{max}}, \forall j \in \mathcal{I}$).

Note that the aforementioned closed access scenario (where $\epsilon_{j} = 0, \forall j \in \mathcal{I}$) is a special case of the hybrid access scenario and the values of $n_{j}^{c}$ (forall $j \in \mathcal{I}$) and $w_{ij}^{c}$ (forall $i \in \mathcal{S}_{j}, j \in \mathcal{I}$) can be determined by applying Algorithm 3.2 at the beginning of the network configuration (before the access control stages). We denote $\mathcal{N}_{\text{FBS}}^{c} = \{1, \ldots, n_{\text{FBS}}^{c}\}$ and $\mathcal{N}_{\text{MBS}}^{c} = \{n_{\text{MBS}}^{c}, \ldots, N\}$ as the set of channels allocated to FBSs and the MBS in the closed access scenario, respectively. Then, in the hybrid access scenario, we only need to determine the additional channels that should be reallocated from the MBS to FBSs who accept to serve MUs, denoted as $N_{a}$, where $N_{a} = \sum_{j \in \mathcal{I} \setminus \{0\}} (n_{j}^{\text{min}} - n_{j}^{c})$ according to (3.18). Then, we allocate these $N_{a}$ channels to FBSs using Algorithm 3.2 by setting $k = n_{\text{FBS}}^{c} + 1$, $n_{\text{FBS}} = n_{\text{FBS}}^{c} + 1$ and $n_{\text{MBS}} = n_{\text{MBS}}^{c} + N_{a}$ in Line 4 and Line 16, respectively. Thus, the desired channel allocation strategy is provided in Algorithm 3.2.
3.4.4 Access Control at FBSs

In this subsection, we study the first two levels of problem $P$, and find the desired access control decisions for FBSs assuming the resource allocation and power management algorithms proposed in previous sections are adopted. We first determine the access ratio $\epsilon$ at FBS $j \in \mathcal{I}\setminus\{0\}$. Obviously, if FBS $j$ does not serve any MU, its access ratio $\epsilon$ should be set to zero. If FBS $j$ serves both FUs and MUs, according to (3.14), given the set of MUs associated with FBS $j$ ($S_m^j$), a smaller access ratio $\epsilon$ may result in a larger number of channels allocated to this FBS ($n_j$), since the channel allocation should satisfy $\epsilon_j n_j \geq \sum_{i \in S_m^j} w_{ij}^{\text{min}}$. Moreover, a smaller access ratio means sharing fewer resources with the MUs. Therefore, in order to maximize the utility, FBS $j$ prefers selecting the minimum achievable access ratio. However, an FBS should also guarantee the minimum QoS requirements for its associated MUs. According to (3.17), the maximum number of channels that FBS $j$ can be guaranteed is $n_c^j + \lfloor \sum_{i \in S_m^j} w_{i0}^c \rfloor$. Thus, the minimum access ratio for FBS $j$ to guarantee its associated MUs’ QoS requirements is

$$\epsilon_j^{\text{min}} = \frac{\sum_{i \in S_m^j} w_{ij}^{\text{min}}}{n_c^j + \lfloor \sum_{i \in S_m^j} w_{i0}^c \rfloor}. \quad (3.21)$$

Note that $n_c^j$ and $w_{i0}^c$ can be calculated by the scheduler and broadcast to the FBSs and MUs at the beginning of the configuration period. In summary, the desired strategy for base station $j \in \mathcal{I}$ in Level II is

$$\epsilon_j^* = \begin{cases} 
0, & \text{if } S_m^j = \emptyset, \\
\epsilon_j^{\text{min}}, & \text{otherwise}. 
\end{cases} \quad (3.22)$$

Finally, we study the user association decision at FBS $j$ when multiple access requests are received from the MUs. As mentioned before, the FBSs are selfish and they select
the desired MUs to maximize their own utilities. For FBS \( j \), the optimal user association decision can be obtained by solving the following problem

\[
\text{maximize}_{a_j} \quad U_j \\
\text{subject to} \quad a_j \in A_j,
\]  

where \( U_j = \sum_{k \in S^f_j} u_{kj}(w^*_k(x^*_j(e^*_j(a_j), a_j)), P_k^*(w^*_k(x^*_j(e^*_j(a_j), a_j)))) \) and \( A_j \) is the set of feasible association profiles defined in Section 3.3.3. \( P_k^*(w^*_k(x^*_j(e^*_j(a_j), a_j))) \), \( w^*_k(x^*_j(e^*_j(a_j), a_j)) \), \( x^*_j(e^*_j(a_j), a_j) \) and \( e^*_j(a_j) \) are obtained using strategies proposed in Sections 3.4.1, 3.4.2, and 3.4.3, and equation (3.22), respectively, for a given association profile \( a_j \). Note that the FBS prefers obtaining more resources from the MBS to serve its FUs. The optimal values of \( a_j \) in (3.23) can be found by searching among the association profiles in \( A_j \) and selecting the one that brings the largest amount of additional resources to the FBS. Specifically, for each association profile, i.e., \( a_j \), the guaranteed number of channels allocated to FBS \( j \) is

\[
n_j = n^c_j + \left\lfloor \sum_{i \in S^M} a_{ij} w^c_{i0} \right\rfloor,
\]

and the additional resources FBS \( j \) obtains by accepting MUs can be represented as

\[
\Delta\varphi_j(a_j) \triangleq \left\lfloor \sum_{i \in S^M} a_{ij} w^c_{i0} \right\rfloor - \sum_{i \in S^M} a_{ij} w^\min_{ij},
\]

where the first term is the additional channels obtained from the MBS and the second term is the amount of resources to be allocated to the MUs. Therefore, the association profile that maximizes \( \Delta\varphi_j(a_j) \) at FBS \( j \) is \( \arg\max_{a_j \in A_j} \Delta\varphi_j(a_j) \). Note that it is not necessary for the FBS \( j \) to obtain more than \( |S_j^f| \) channels to serve its FUs, since we assume each FU can only access one channel at a time. We denote \( A'_j = \{ a_j \mid a_j \in A_j, n^c_j + \Delta\varphi_j(a_j) > |S_j^f| \} \).
Chapter 3. Network Configuration for Two-tier Heterogeneous Networks

It can be seen that if $A_j' \neq \emptyset$, accepting any association profile in $A_j'$ maximizes the total utility of FUs at FBS $j$, since the resources available for the FUs are sufficient to guarantee that each FU obtains one channel. In this case, we choose the profile that contains the minimum number of MUs as the association decision. On the other hand, if $A_j' = \emptyset$, the optimal association decision can be selected as the association profile that gives the largest $\Delta \varphi_j(a_j)$. Therefore, the optimal user association decision for FBS $j$ can be represented as

$$a_j^* = \begin{cases} 
\arg \min_{a_j \in A_j'} \sum_{i \in S^M} a_{ij}, & \text{if } A_j' \neq \emptyset, \\
\arg \max_{a_j \in A_j} \Delta \varphi_j(a_j), & \text{otherwise.}
\end{cases}$$

(3.26)

3.4.5 Procedures of the Proposed Mechanism

We have derived the desired strategies for base stations and users in each network configuration stage. The proposed network configuration mechanism can be summarized as follows.

(1) **Initialization**: The scheduler collects global information of the network, and calculates the resource allocated to the FBSs and MUs using Algorithms 3.2 and 3.1, respectively, assuming closed access is used at all FBSs. Then, it sends the information $n_c^j$ and $w_{c0}^i, \forall i \in S^M$ to each FBS $j \in I\{0\}$. The scheduler also obtains the channel allocation decisions in the closed access mode, where the sets of channels allocated to the FBSs and the MBS are $\{1, \ldots, n_c^FBS\}$ and $\{n_c^{MBS}, \ldots, N\}$, respectively.

(2) **Access Control Process**: Each MU sends an access request to the FBS from which the MU can obtain the largest utility. Then, each FBS $j$ determines its association decision according to equations (3.26) and (3.22), and sends the information of $a_j$ and $\epsilon_j$ to the scheduler. If an MU is not selected by any FBS, it will be served by the MBS.

(3) **Resource Allocation Process**: The scheduler first determines the number of additional channels ($N_a$) to be reallocated from MBS to FBSs based on the resource allocation
decisions in closed access mode (obtained in the initialization step) and the access control decisions. Next, the scheduler calculates the minimum number of channels required at each base stations according to (3.18) and allocates the additional channels \(N_a\) one by one using Algorithm 3.2 by setting \(k = n_{\text{FBS}}^c + 1\), \(n_{\text{FBS}} = n_{\text{FBS}}^c + 1\) and \(n_{\text{MBS}} = n_{\text{MBS}}^c + N_a\) in Line 4 and Line 16, respectively. Then, the MBS and FBSs allocate their available resources to each associated user according to the results obtained using Algorithm 3.1 during the channel allocation process.

(4) Power Management Process: Each user determines its optimal power according to equation (3.10) based on the resources allocated by its associated base station.

Note that the proposed mechanism explores channel reuse among femtocells, which significantly improves the system performance. Moreover, the proposed mechanism has a nice property, it encourages an FBS to use hybrid access control in order to improve the total utility of its FUs. On one hand, an FBS may obtain additional resources to serve its own FUs by accepting to serve nearby MUs, as discussed in Section 3.4.4. On the other hand, an MU who communicates with the MBS with poor signal quality can improve its utility by accessing the nearby FBS without additional cost (i.e., extra payment). This type of incentive is important since our model considers FBSs are installed by selfish private users who may not share their services for free. Without such incentive, FBSs may refuse to serve MUs and always use closed access mode.

3.4.6 Complexity of the Proposed Algorithms

In the proposed configuration mechanism, the optimization in Levels II and V takes constant time, as can be seen from (3.22) and (3.10), respectively. From (3.26), the computational complexity of Level I is proportional to the cardinality of the feasible set \(|A_j|\). Therefore, the major computation complexity originates from Algorithms 3.1 and
3.2. Both algorithms use iterative approaches. In the following, we provide an analysis of these two algorithms with respect to the number of iterations. In Algorithm 3.1, the basic allocation step from Line 2 to Line 3 can be computed in constant time. Therefore, the complexity mainly depends on the loop from Line 7 to Line 13. Note that in each iteration, we allocate one time slot to a user. There are \( T_R = T(n_j - \sum_{i \in S_j} w_{ij}) \) iterations in this loop. Within each iteration, Line 10 searches for the user with the largest utility increment, which can be implemented using binary search. The average number of searching iterations for this step is \( \log(|S_j|) \). Therefore, the complexity of Algorithm 3.1 with respect to the number of iterations is \( T_R \log(|S_j|) \), and each iteration takes constant time for computation. Algorithm 3.2 also contains two steps, and the major complexity depends on the second step from Line 17 to Line 33. We denote the number of remaining channels after the basic allocation as \( N_R \). In the second step, we allocate the channels to each base station one by one iteratively. For simplicity, we assume all FBSs have the same number of subscribed FUs. In this case, the maximum number of iterations for the outer loop is \( N_R \). Within each iteration for channel allocation, we check all base stations one by one. Each time we need to invoke Algorithm 3.1 to update the utility increment with one additional channel, which further takes \( T \log(|S_j|) \) iterations. Thus, the complexity with respect to the number of iterations in this worst case is \( N_R(|I|)T \log(|S_j|) \). It can be seen that the complexity increases linearly with the number of remaining channels, the number of FBSs, and the number of time slots in the transmission phase.

3.5 Performance Evaluation

In this section, we evaluate the performance of the proposed network configuration mechanism using simulation. We consider a system that consists of an MBS and 10 FBSs.
which are close to each other. The FBSs are randomly located within a circular region whose radius is $r$. The distance between the MBS and the region center is $d$. The radius of each femtocell is 10 m, and four FUs are randomly distributed within each femtocell. The MBS serves 15 MUs who are randomly distributed in the region with the constraint that at least 8 of them are indoor users who are located within at least one femtocell. The maximum number of users an FBS can serve is $L = 8$. There are $N$ channels available for the system, each with a bandwidth of 180 kHz. The wireless channel model follows [88], where the path loss exponent between a user and the MBS (or an FBS) is 4 (or 3), and we choose the wall penetration loss as 8 dB. The users’ maximum transmission power is $P_{\text{max}} = 250$ mW and the noise power is $-120$ dBm. The interference threshold $\theta$ is $-90$ dBm. Each time slot is 1 ms and one transmission period consists of $T = 1000$ time slots. We consider two types of applications, the MUs are requesting regular video chatting with the minimum rate requirement ($R_{\text{min}}^i$) of 256 kbps. The FUs are requesting high definition video chatting with minimum rate requirement of 400 kbps. Similar to [85], we implement the user satisfaction function with $\Phi_i(\tilde{R}_{ij}) = \alpha \ln(1 + (\tilde{R}_{ij} - R_{\text{min}}^i)) + C$, where $\alpha = 200$ and $C = 50$. We choose $\beta = 0.1$. In this section, the utility of a user is calculated using the actual average data rate, which is obtained by averaging the instantaneous data rate over the transmission phase (1000 time slots). We adopt a random scheduling algorithm to determine the set of time slots for each user based on the resource sharing variables. Specifically, at each time slot, FBS $j$ randomly selects $n_j$ users from the scheduling set, which initially contains all the users to be served. Once the number of time slots allocated to a user reaches the desired number, the user is removed from the scheduling set. We simulate the proposed configuration mechanism and compare its performance with configuration mechanism with closed access and configuration mechanism with orthogonal channel allocation and
Figure 3.4: Total utility of the system versus total number of channels $N$, with $r = 40$ m and $d = 150$ m.

topology-based access introduced in [78], respectively.

We first evaluate the performance of the aforementioned configuration mechanisms with respect to different number of channels $N$, and the results are averaged over 50 simulation runs with $r = 40$ m and $d = 150$ m. Fig. 3.4 shows the total system utility achieved by different network configuration mechanisms, and Fig. 3.5 shows their corresponding system throughput. It can be seen that the proposed mechanism achieves the largest total utility and the highest throughput among the three mechanisms, and the performance gap between the proposed mechanism and the other two mechanisms becomes smaller as the number of channels increases. The reason is as follows: Using the proposed mechanism, when the number of channels is small, these MUs who receive low service quality may switch to a nearby femtocell to improve their utility, which may also improve the total utility (or throughput) of the FUs in that cell. Similarly, the orthogonal channel allocation mechanism with topology-based access can also improve the system utility (or throughput) by accepting MUs at the FBSs. However, this approach allocates orthogonal channels to
each FBS, which neglects the benefit of channel reuse among the FBSs. In addition, since
the access control is purely based on topology information, it is possible that an FBS may
reject the requests of some MUs and miss the opportunity to improve its own performance.
It is also possible that an FBS accepts to serve some MUs but does not obtain additional
channels from the MBS, which may degrade the utility of FUs due to resource sharing.
Therefore, the performance of the mechanism with topology-based access control is not
as good as the proposed mechanism. As the number of channels increases, some FBSs or
MUs may obtain sufficient resources, and access control between these MUs and FBSs are
not necessary. Thus, the performance gap between the proposed mechanism and the other
mechanisms becomes smaller.

In Table 3.2, we show the average running time of the proposed mechanism and the
orthogonal channel allocation mechanism with respect to different number of channels using
MATLAB. As the number of channels increases, the average running time of these two
mechanisms increases, since the computational complexity increases due to the increased
Table 3.2: Average running time versus number of channels

<table>
<thead>
<tr>
<th>$N$</th>
<th>26</th>
<th>28</th>
<th>30</th>
<th>32</th>
<th>34</th>
<th>36</th>
<th>38</th>
<th>40</th>
</tr>
</thead>
<tbody>
<tr>
<td>Proposed mechanism</td>
<td>6.33 s</td>
<td>6.36 s</td>
<td>6.95 s</td>
<td>7.09 s</td>
<td>7.13 s</td>
<td>7.23 s</td>
<td>7.48 s</td>
<td>7.63 s</td>
</tr>
<tr>
<td>Orthogonal channel allocation</td>
<td>9.00 s</td>
<td>10.22 s</td>
<td>10.93 s</td>
<td>11.33 s</td>
<td>11.62 s</td>
<td>11.87 s</td>
<td>12.13 s</td>
<td>12.70 s</td>
</tr>
</tbody>
</table>

Figure 3.6: Total utility of the system versus radius of the region, with $N = 30$, and $d = 150$ m.

Number of variables. It is implied in Table 3.2 that these two mechanisms have similar computational complexity. However, the proposed mechanism achieves better total utility and system throughput, as shown in Figs. 3.4 and 3.5.

Next, we evaluate the performance of the three mechanisms with respect to different radii of the circular area $r$. It is shown in Fig. 3.6 that the system utility of the proposed mechanism and the closed access mechanism increase as the radius of the region increases. This is because as the radius increases, the average distance between two randomly distributed FBSs becomes larger, which increases the possibility of channel reuse among the FBSs. The orthogonal channel allocation mechanism remains almost constant as $r$ changes. This is because this mechanism uses an orthogonal channel
allocation scheme without the consideration of interference among the femtocells, and the decision is not affected by the radius of the circular region. Therefore, when the radius is small, the proposed mechanism with hybrid access achieves a similar performance to the orthogonal channel allocation mechanism since channel reuse rarely happens due to the close proximity of the FBSs, as the radius increases, the proposed mechanism achieves significant utility improvement and the performance gap between the two mechanisms becomes larger.

Then, we adjust the distance between the MBS and the considered region center, and evaluate the performance of the configuration mechanisms with respect to different values of $d$. Fig. 3.7 shows that the system utility decreases as $d$ increases for the proposed mechanism and the closed access mechanism. However, the utility decrement step size of the proposed mechanism becomes smaller when $d$ increases from 100 m to 140 m and then increases thereafter. This is because when $d$ is small, the signal quality from the MBS to the MUs are good enough to satisfy their QoS requirements and accessing the MBS may achieve a larger utility for each MU. As $d$ increases, the signal quality from the MBS to
Chapter 3. Network Configuration for Two-tier Heterogeneous Networks

Figure 3.8: Total utility versus number of MUs moved into the femtocell region with $N = 30$, $d = 150$ m and $r = 40$ m.

the MUs degrades accordingly, which decreases the total utility. When $d$ is greater than a certain value, i.e., 100 m, the signal quality from the MBS to the MUs is poor and the MUs may request to access their nearby FBSs to improve their utilities. With the proposed mechanism, the FBSs can accept to serve the MUs who experience severe channel fading from the MBS, which improves both the utilities of MUs and FUs when $d$ varies from 100 m to 140 m. After that, FBSs may not accept more MUs, and the total utility decrease linearly as $d$ increases. A similar trend can be seen for the orthogonal channel allocation mechanism.

In Fig. 3.8, we adjust the locations of MUs and evaluate their impact on the performance of the proposed mechanism. We randomly position MUs in a ring area and the average distance between MU to the region center is 80 m. We then move a number of MUs into the circular region where the FBSs are located. It is shown in Fig. 3.8 that the system utility of the proposed mechanism first increases and then decreases as more MUs move into femtocells, and the performance gap between the proposed mechanism
and the closed access mechanism becomes larger. The reason is that as the MUs move into the femtocells, their minimum resource demand from the MBS becomes larger due to the wall penetration loss of signal, which decreases their utility in the closed access mode. However, as MUs move closer to the FBSs, they require less resources from the FBSs to satisfy their minimum rate requirements. In this case, the proposed mechanism improves the performance of MUs and FUs by redistributing the channels among the MBS and FBSs, which results in an increase of the performance gap. As more MUs moves into the femtocells, once the utility improvement with hybrid access does not compensate for the utility decrease (for these MUs), the system utility starts to decrease. Nevertheless, the proposed mechanism achieves superior performance than the mechanism with closed access mode.

Finally, we evaluate the performance of the proposed mechanism with respect to different values of the weighting factor $\beta$. We select $\beta \leq C/P_{\text{max}}$ to guarantee positive utility of each user in the system. We also implement another user satisfaction function.
[89] as \( \Phi_i^j(\tilde{R}_{ij}) = \alpha(1 - \exp(- (\tilde{R}_{ij} - R_{i}^{\min})/R_{i}^{\min})) + C. \) Fig. 3.9 shows the users’ satisfaction and the energy consumption decrease as \( \beta \) increases for both scenarios. The reason is that when \( \beta \) is small, improving users’ satisfaction with respect to data rate is more important than saving energy, and users may use large transmission power to achieve high data rates. However, as \( \beta \) increases, the energy consumption component in the utility function becomes more critical. To maximize the utility, users may consider reducing transmission power to save energy. Therefore, we can adjust the parameter \( \beta \) accordingly to balance the trade-off between users’ satisfaction and energy consumption, and to realize different optimization objectives. Note that this result is valid for other types of concave satisfaction functions as well, since the proposed mechanism is shown to achieve optimal resource allocation solution with any concave satisfaction function.

### 3.6 Summary

In this chapter, a multi-stage network configuration mechanism was proposed for two-tier macro-femto heterogeneous networks, which consists of access control, channel allocation and power management processes. This mechanism employed a resource allocation approach to motivate hybrid access at the FBSs. The configuration mechanism was modeled as a multi-stage decision making process, where the scheduler, base stations and users make decisions sequentially to optimize their own utilities. Based on a multi-level optimization approach, efficient algorithms were proposed to find the desired decision in each network configuration process. Simulation results showed that the proposed network configuration mechanism achieved a higher system utility than mechanism with topology-based hybrid access or mechanism with closed access, especially when the total number of available channels is small.
Chapter 4

Resource Sharing for Cloud-based Radio Access Networks

4.1 Introduction

This chapter aims to develop a resource sharing mechanism to support multiple service providers in a C-RAN. As introduced in Chapter 1, due to the centralization of BBUs and the advance of cloud computing technology, sharing C-RAN resources among multiple providers has become a promising approach to reduce CAPEX and OPEX for each service provider. However, designing resource sharing mechanism in C-RAN is challenging due to interference, user mobility, and traffic demand variation. Specifically, managing interference usually requires coordination or joint resource allocation among multiple RRHs. Moreover, when users are mobile and their traffic demand varies frequently, the resources reserved for a service provider may not be sufficient to satisfy the QoS of its users. In this chapter, these issues are studied and a multi-timescale resource sharing mechanism is proposed to dynamically allocate radio resources while providing service isolation among the service providers. A new metric is proposed to determine the minimum aggregate data rate guaranteed for each service provider based on its users’ QoS requirements and channel conditions. A threshold-based policy is applied to control the maximum interference among different service providers in order to guarantee service isolation. To handle traffic demand variation, this mechanism performs
Chapter 4. Resource Sharing for Cloud-based Radio Access Networks

a global optimization in a large timescale and several local resource optimizations in a relatively small timescale. Moreover, a mobility prediction approach is employed to address the user mobility issue.

The rest of this chapter is organized as follows. A summary of related works is provided in Section 4.2. In Section 4.3, we describe the system model and the user-centric network sharing scheme, and formulate the resource allocation problem. In Section 4.4, we propose an efficient algorithm to solve the resource allocation problem. Section 4.5 describes the proposed multi-timescale resource sharing mechanism. Performance of the proposed mechanism is evaluated in Section 4.7. Summary is provided in Section 4.8.

4.2 Related Works

Resource sharing in cellular networks has been studied during the past few years. Some researchers have proposed base station sharing mechanisms [90, 91]. Kokku et al. in [90] propose a network virtualization substrate, where a slice scheduler is integrated into the base station’s scheduling component for managing the resource slicing and sharing among different service providers. This approach provides flow level isolation and customization by partitioning the available channels into non-overlapping slices. Similar approaches have been applied in [91] under a different network setting. Since these approaches focus on resource virtualization in a single base station, the effect of interference among different base stations has not been considered. Therefore, they may not be suitable for densely deployed small cell networks such as C-RAN. Other network sharing mechanisms are developed based on dynamic allocation of the spectrum resources [92–94]. In these approaches, the physical infrastructure can be used by all service providers, and service isolation is provided by allocating different spectrum to each service provider. Fu et al. in [92] map the wireless spectrum resources into a rate region, and propose a sequential
Chapter 4. Resource Sharing for Cloud-based Radio Access Networks

auction game framework to allocate the resources to several competing service providers. In [93], an opportunistic resource sharing scheme is proposed, which explores the varying traffic patterns and allows different traffic flows to access the same channel opportunistically. In [94], Hou et al. formulate the spectrum sharing problem for multi-hop software defined radio networks as a mixed-integer non-linear programming problem, and design an algorithm to find the near optimal solution. Although these approaches guarantee service isolation, they exclude the possibility of spectrum reuse among service providers. Resource sharing in cloud-based networks has also been studied [95, 96]. A number of sharing policies, such as network proportionality, have been proposed and evaluated in [95]. Efficient bandwidth reservation with varying traffic demand is studied in [96] for resource sharing in data centers. However, these works do not consider sharing the wireless resources, which cannot be directly applied to C-RANs.

4.3 System Model

We consider a C-RAN as shown in Fig. 4.1, which consists of RAN and a cloud-based data center. The RAN consists of a set of small cell RRHs (i.e., microcell or picocell RRHs), which is denoted as $\mathcal{B}$. All RRHs are connected to the data center via high-speed optical fiber. We assume the capacity of each optical fiber link is large enough to handle the data transmission between the data center and the corresponding RRH. The data center is responsible for performing resource management and control operation for the RAN. The available spectrum for this system is divided into $N$ orthogonal channels with equal bandwidth $B$. Let $\mathcal{N} = \{1, \ldots, N\}$ denote the set of channels. We consider path loss and shadowing effect of the wireless channel, and the average channel gain between a user and a RRH is distance-dependent. The system is time-slotted, where only one user can access a particular channel from a RRH during one time slot. However, users may share the same
Figure 4.1: A cloud-based radio access network. Multiple service providers lease the infrastructure and radio resource from a network operator. The network operator creates vRAN for each service provider by assigning a number of RRHs and the corresponding channel resources. Each RRH can serve users from different service providers.

channel via time division multiple access. The same channel can be reused by multiple RRHs to improve the system capacity under certain interference constraints specified by the network operator.

We consider a number of service providers share the C-RAN. The C-RAN operator owns the infrastructure and the spectrum, and can lease these resources to a set of service providers, denoted as \( S \). Each service provider \( s \in S \) provides data services with certain QoS requirements, such as video streaming and sports TV broadcasting, to a set of subscribed users, which is denoted as \( U_s \). To enable resource sharing among service providers, the network operator employs virtualization techniques to create vRANs and allocate corresponding resources periodically. Specifically, at the beginning of the virtualization process, each service provider sends a reservation request for certain traffic demand (or aggregate data rate required from the users) to the network operator. Then, the network operator creates a vRAN for each service provider, which consists of a
number of vRRHs and BBUs in the data center as shown in Fig. 4.1. Each vRRH can be mapped to a real RRH in the system. Multiple vRRHs from different service providers can be mapped to the same RRH and they share the resources available at that RRH. The network operator also determines the amount of channel resources allocated to each vRRH. After that, each service provider performs scheduling and data transmission in its corresponding vRAN. The vRAN created for each service provider remains unchanged until the next resource allocation is performed at the network operator. This process is performed every $T$ time slots, and we refer $T$ time slots as a resource sharing period.

### 4.3.1 User-Centric Resource Sharing Scheme

In conventional resource sharing schemes, the service providers need to estimate their users’ resource demand before sending the reservation requests. Performing such estimation usually requires the knowledge of user association decisions to estimate the channel gain for each user and the available channel information at each base station to estimate the corresponding interference. Most of the existing estimation approaches assume simple resource allocation and user association decisions, i.e., the channels available at each base station are either orthogonally allocated or follow a fixed reuse pattern. This may lead to unbalanced resource reservation in the network and competition among the service providers when the users are not uniformly distributed. In this chapter, considering the cloud computing capability in C-RANs, we shift the resource demand estimation task from the service providers to the network operator, and propose a user-centric resource sharing scheme. The basic idea is to let the service providers send the QoS requirements for their subscribed users to the network operator. Based on the users’ information, the network operator performs joint admission control, user association, and resource allocation to create the vRANs with guaranteed QoS of
admitted users. Different from the existing resource sharing schemes, we propose to use
the following metrics and approaches during the network-wide sharing process:

1) *Interference threshold*: Due to channel reuse among the RRHs, different scheduling
and transmission decisions at one service provider may affect the level of interference
experienced in other vRANs, which further affect other service providers’ decisions. To
tackle this problem, we introduce an interference threshold $\xi$ to assist the service
isolation process. Specifically, during resource allocation, we require that the interference
experienced at each user in a vRAN should not exceed this threshold. Thus, the impact
of interference among vRANs is limited within a controllable range. A service provider
can make scheduling decisions based on the interference threshold instead of the actual
interference originated from other vRANs. The value of the threshold can be selected as
the one that maximizes the average system throughput, which can be obtained via
computer simulations.

2) *Rate estimation with mobility prediction*: The resource demand estimation is
performed by the network operator at the beginning of a resource sharing period, i.e., at
time slot $t$ in period $[t, t+T)$. When a user is mobile, the results obtained based on the
information (e.g., channel gain) at time slot $t$ may not be sufficient to guarantee its QoS
during the entire period $[t, t+T)$. To address this issue, we exploit the users’ mobility
information when estimating their achievable data rates. Specifically, at time slot $t$, we
consider the network operator can predict the locations of the user at future time slots
$t + \Delta t, t + 2\Delta t, \ldots, t + T - \Delta t$ according to a certain mobility prediction mechanism [97],
where $\Delta t = T/n_p$ and $n_p$ is a pre-determined positive integer that represents the number
of locations we used for estimation. We denote the predicted location at time slot $t + \tau \Delta t$
as the $\tau$th predicated location, where $\tau$ takes values from set $\mathcal{T} = \{0, 1, \ldots, n_p - 1\}$. For
user $u \in \mathcal{U}_s$, we define $\mathcal{B}_{u,\tau}, \forall \tau \in \mathcal{T}$ as the possible set of RRHs to be associated at the
Chapter 4. Resource Sharing for Cloud-based Radio Access Networks

\( \tau \)th predicted location. We denote \( a_{u,j,\tau} \in \{0, 1\} \) as the corresponding association variable where \( a_{u,j,\tau} = 1 \) indicates user \( u \) is served by RRH \( j \in \mathcal{B}_{u,\tau} \) at the \( \tau \)th predicted location, and \( a_{u,j,\tau} = 0 \) otherwise. We further denote \( w_{u,j,\tau} \in [0, 1/n_p], \forall j \in \mathcal{B}_{u,\tau}, \tau \in \mathcal{T} \) as the resource sharing variable, which indicates the proportion of time that user \( i \) can access a channel at RRH \( j \) during the period of \( T \) time slots. The variables \( a_{u,j,\tau} \) and \( w_{u,j,\tau} \) remain unchanged within \( \Delta t \) time slots. Then, the estimated data rate for a user \( u \in \mathcal{U}_s \) during one resource sharing period is

\[
R_u = \sum_{j \in \mathcal{B}_{u,\tau}, \tau \in \mathcal{T}} w_{u,j,\tau} B \log_2 \left( 1 + \frac{P_{RRH} g_{u,j,\tau}}{\sigma^2 + \xi} \right), \quad (4.1)
\]

where \( g_{u,j,\tau} \) is the channel gain between user \( u \) and RRH \( j \) at the \( \tau \)th predicted location, \( P_{RRH} \) is the transmission power of the RRH, and \( \sigma^2 \) is the noise power. Note that the definition in (4.1) involves the interference threshold \( \xi \) and predicted channel gains \( g_{u,j,\tau} \).

The actual data rate of a user in a particular time slot depends on the actual channel gain and the aggregate interference. The estimated data rate is used for resource allocation at time slot \( t \) since the network operator does not know exactly the channel gains in future time slots.

3) Dynamic rate guarantee: The network operator should preserve some resources for each service provider to guarantee a minimum throughput (or aggregate data rate of users) at the service provider. However, since users are mobile, the amount of radio resources consumed for guaranteeing a fixed aggregate data rate for a service provider varies, which may result in unbalanced resource allocation and low system throughput. For example, a service provider may consume a large amount of wireless resources when the subscribed users experience poor channel conditions. Thus, the minimum aggregate data rate guaranteed for each service provider should be dynamically adjusted in order to improve the system throughput. We propose the following metric to determine the minimum aggregate data
rate guaranteed for each service provider. We define $R_{s\text{ref}}$ as the reference minimum aggregate data rate for service provider $s \in S$, and $R_{ref}$ as the reference data rate for a user. $R_{s\text{ref}}$ is the upper bound of the minimum rate guaranteed by the network operator, which is specified in the service agreement. $R_{\text{ref}}$ is calculated as the data rate of a user at a reference distance (i.e., 20 m) to a RRH. We denote the maximum achievable data rate of user $u \in U_s$ as

$$R_u^* = \frac{1}{n_p} \sum_{\tau \in \mathcal{T}} B \log_2 \left( 1 + \frac{P_{\text{RRH}} g_{u,j^*,\tau}}{\sigma^2 + \xi} \right),$$  \hfill (4.2)$$

where $g_{u,j^*,\tau}$ is the channel gain between user $u$ and its closest RRH $j^*$ at the $\tau$th predicted location. Then, the minimum aggregate data rate guaranteed for service provider $s$, denoted as $R_{s\text{min}}$, is determined according to the following rules: (i) $R_{s\text{min}}$ should be no larger than the reference value $R_{s\text{ref}}$. (ii) $R_{s\text{min}}$ should be no larger than the maximum traffic demand from the subscribed users, which is $\sum_{u \in U_s} R_{u\text{max}}$. (iii) When the average value of users’ maximum achievable data rates, $(1/|U_s|)\sum_{u \in U_s} R_u^*$, is smaller than the reference data rate $R_{\text{ref}}$, it implies that on average the users are relatively far from their closest RRH and guaranteeing $R_{s\text{ref}}$ consumes more resources than expected. In this case, the network operator only guarantees a lower data rate (downscale $R_{s\text{ref}}$ by a factor of $(1/|U_s|)\sum_{u \in U_s} R_u^*/R_{\text{ref}}$) for the service provider to save some resources. In summary, we have

$$R_{s\text{min}} = \min \left\{ R_{s\text{ref}}, \sum_{u \in U_s} R_{u\text{max}}, \frac{\sum_{u \in U_s} R_u^*}{|U_s| R_{\text{ref}}}, R_{s\text{ref}} \right\},$$  \hfill (4.3)$$

4) Balanced optimization objective: We consider the network operator jointly maximizes the number of admitted users and the system throughput under different traffic load situations,
which can be characterized using the following objective function

$$f = \sum_{u \in \mathcal{U}, s \in \mathcal{S}} \alpha_u z_u + \sum_{u \in \mathcal{U}, s \in \mathcal{S}} \beta_u R_u,$$

(4.4)

where $z_u \in \{0, 1\}$ is the admission control variable for user $u$, and $\alpha_u, \beta_u, \forall u \in \mathcal{U}, s \in \mathcal{S}$ are time-dependent and user-dependent weighting factors. The first term on the right hand side (r.h.s.) of (4.4) is the weighted total number of admitted users in the system during this allocation period. The second term on the r.h.s. of (4.4) characterizes the weighted sum rate of the system, which is equivalent to the system throughput when $\beta_u = 1, \forall u \in \mathcal{U}, s \in \mathcal{S}$. The weighting factors $(\alpha_u, \beta_u)$ are used to characterize the importance of the two objectives. $\alpha_u$ can also be interpreted as a priority factor for user $u$ during the admission control process, where a larger $\alpha_u$ indicates a higher priority. This objective function can be used to characterize different optimization objectives by adjusting the weighting factors. For example, if $\alpha_u = 1, \beta_u = 0, \forall u \in \mathcal{U}, s \in \mathcal{S}$, the operator targets guaranteeing service for as many users as possible. When $\alpha_u = 0, \beta_u = 1, \forall u \in \mathcal{U}, s \in \mathcal{S}$, the operator aims to maximize the system throughput.

Based on the previous discussion, in the proposed scheme, the network operator determines the number of channels allocated to each RRH, the amount of resources shared by each service provider, as well as the admission control and user association decisions.

### 4.3.2 Resource Allocation Problem at the Operator

We consider resource allocation at the beginning of a $T$-time-slot period $[t, t + T)$, and define $C_{j,k} \in \{0, 1\}$ as the channel allocation variable, where $C_{j,k} = 1$ indicates channel $k \in \mathcal{N}$ is allocated to RRH $j \in \mathcal{B}$, and $C_{j,k} = 0$ otherwise. We let $C = (C_{j,k}, j \in \mathcal{B}, k \in \mathcal{N})$. 
We let $z_s = (z_u, u \in U_s)$. The user association variables and resource sharing variables for service provider $s \in S$ can be represented as $a_s = (a_{u,j,\tau}, u \in U_s, j \in B_{u,\tau}, \tau \in T)$ and $w_s = (w_{u,j,\tau}, u \in U_s, j \in B_{u,\tau}, \tau \in T)$, respectively. Then, the resource allocation problem is to determine the decision variables $z_s, a_s, w_s, \forall s \in S,$ and $C$, which can be formulated as the following optimization problem.

$$\begin{align*}
\text{maximize} & \quad \sum_{u \in U_s, s \in S} \alpha_u z_u + \sum_{u \in U_s, s \in S} \beta_u R_u \\
\text{subject to} & \quad z_u R_{u \text{min}} \leq R_u \leq z_u R_{u \text{max}}, \quad \forall u \in U_s, s \in S, \\
& \quad \sum_{u \in U_s, \tau} w_{u,j,\tau} \leq \frac{1}{n_p} \sum_{k \in N} C_{j,k}, \quad \forall j \in B, \tau \in T, \\
& \quad a_{u,j,\tau} \sum_{i \in B \setminus \{j\}} C_{i,k} P_{\text{RRH}u,i,\tau} \leq \xi, \quad \forall k \in N, j \in B, \tau \in T, u \in U_s, s \in S, \\
& \quad \sum_{u \in U_s} R_u \geq R_{s \text{min}}, \quad \forall s \in S, \\
& \quad \sum_{j \in B_{u,\tau,\tau}} w_{u,j,\tau} \leq 1, \quad \forall u \in U_s, s \in S, \\
& \quad a_{u,j,\tau} \frac{1}{T} \leq w_{u,j,\tau} \leq a_{u,j,\tau} \frac{1}{n_p}, \quad \forall j \in B_{u,\tau,\tau} \in T, u \in U_s, s \in S, \\
& \quad C_{j,k} \in \{0,1\}, \quad \forall j \in B, k \in N, \\
& \quad z_u \in \{0,1\}, \quad \forall u \in U_s, s \in S, \\
& \quad a_{u,j,\tau} \in \{0,1\}, \quad \forall j \in B_{u,\tau,\tau} \in T, u \in U_s, s \in S.
\end{align*}$$

Constraint (4.5b) guarantees the achievable data rate for each admitted user $u$ be within the range $[R_{u \text{min}}, R_{u \text{max}}]$. Constraint (4.5c) implies that during any time period $[t+\tau \Delta t, t+\tau (\tau + 1) \Delta t)$, $\forall \tau \in T$, the total resources allocated to users from RRH $j$ should not exceed its total available resources, where $\tilde{U}_{j,\tau} \triangleq \{u \mid j \in B_{u,\tau}, \tau \in T, u \in U_s, s \in S\}$ is the set of users that can be associated with RRH $j$. Constraint (4.5d) indicates that the aggregate
interference at any user $u$ associated with RRH $j$ over its allocated channel $k$ should not exceed the threshold value $\xi$. The minimum resource guarantee for each service provider is specified in (4.5e). Constraints (4.5f) and (4.5g) state the range of the resource sharing variable $w_{u,j,\tau}$ and its relationship with the user association variable $a_{u,j,\tau}$. Constraint (4.5f) implies a user can only access one channel at a time. Note that the minimum resource unit is one time slot, and the maximum number of time slots available for each user in $[t + \tau \Delta t, t + (\tau + 1) \Delta t]$ is $\Delta t = T/n_p$. Thus, we have $1/T \leq w_{u,j,\tau} \leq \Delta t/T = 1/n_p$ for user $u$ served by RRH $j$, which is implied in constraint (4.5g).

Note that constraint (4.5d) is non-convex, problem (4.5) is a mixed-integer non-linear problem (MINLP) with non-convex constraints, which is difficult to solve in practice. In the following sections, we design an efficient algorithm to find a suboptimal solution.

### 4.4 Efficient Resource Allocation Algorithm

#### 4.4.1 Problem Transformation

In this section, we transform problem (4.5) into a mixed-integer linear programming (MILP). Specifically, we convert the non-convex constraint (4.5d) into the following linear constraint

$$\sum_{l \in B \setminus \{j\}} C_{l,k} P_{\text{RRH}u,l,\tau} \leq \frac{(a_{u,j,\tau} + C_{j,k})\xi}{2} + (2 - a_{u,j,\tau} - C_{j,k})D,$$

$$\forall k \in \mathcal{N}, j \in \mathcal{B}, \tau \in \mathcal{T}, u \in \mathcal{U_s}, s \in \mathcal{S},$$

where $D$ is a large constant. Regarding this linearization, we have the following theorem.
Theorem 4.1 Constraint (4.6) is equivalent to constraint (4.5d) if \( D \) and \( \xi \) satisfy

\[
D \geq \max \left\{ \frac{\Gamma}{2}, \frac{\Gamma}{2} \right\},
\]

where

\[
\Gamma = \left( |\mathcal{B}| - 1 \right) P_{\text{RRH}} \max_{u \in \mathcal{U}, s \in \mathcal{S}, j \in \mathcal{B}, \tau \in \mathcal{T}} \{ g_{u,j,\tau} \}.
\]

Proof We show constraints (4.5d) and (4.6) are equivalent under condition (4.7) when \( a_{u,j,\tau} \) and \( C_{j,k} \) take any feasible values from \{0, 1\}. First, when \( a_{u,j,\tau}C_{j,k} = 1 \) (or \( a_{u,j,\tau} = 1 \) and \( C_{j,k} = 1 \)), constraints (4.5d) and (4.6) are the same. Second, when \( a_{u,j,\tau}C_{j,k} = 0 \), (4.5d) is always satisfied. Thus, we only need to show that (4.6) is always satisfied when \( a_{u,j,\tau}C_{j,k} = 0 \). We consider the following cases for \( a_{u,j,\tau}C_{j,k} = 0 \): When \( a_{u,j,\tau} = 0 \) and \( C_{j,k} = 0 \), constraint (4.6) becomes \( \sum_{l \in \mathcal{B} \setminus \{j\}} C_{l,k} P_{\text{RRH}} g_{u,l,\tau} \leq 2D \). When \( a_{u,j,\tau} = 1 \) and \( C_{j,k} = 0 \) or \( a_{u,j,\tau} = 0 \) and \( C_{j,k} = 1 \), constraint (4.6) becomes \( \sum_{l \in \mathcal{B} \setminus \{j\}} C_{l,k} P_{\text{RRH}} g_{u,l,\tau} \leq \frac{\xi}{2} + D \). Based on (4.8), we have \( \Gamma \geq \sum_{l \in \mathcal{B} \setminus \{j\}} C_{l,k} P_{\text{RRH}} g_{u,l,\tau} \). Thus, according to condition (4.7), it can be verified that (4.6) is always satisfied in all three cases. This completes the proof.

With the linear constraint (4.6), problem (4.5) becomes

\[
\begin{align*}
\text{maximize} \quad & f \\
\text{subject to} \quad & (4.5b), (4.5c), (4.5e)–(4.5j), \text{ and } (4.6). 
\end{align*}
\]

Problem (4.9) is an MILP, which can be solved by applying standard techniques such as the branch and bound method. However, the computational complexity of these techniques increases significantly as the size of the problem increases. Therefore, in a system with many users, finding the optimal solution to problem (4.9) by applying the standard techniques may consume a large amount of time, which may not be practical for
real-time processing. To address this issue, in the next subsection, we propose a fast algorithm to find an efficient suboptimal solution.

### 4.4.2 Increment-based Greedy Allocation Algorithm

To jointly determine the channel allocation, resource sharing, and user admission and association decisions, we propose an increment-based greedy allocation (IBGA) algorithm. The basic idea of the proposed algorithm is to allocate the available channels to the RRHs one by one. For each channel, we allocate it to the RRHs iteratively, where in each iteration we select the RRH that has the largest increment of the objective value while satisfying the interference constraint (4.6). The allocation of a channel terminates when no more RRH can use this channel under the interference constraint. Once the channel allocation is fixed, the user admission and association are also determined accordingly. To characterize the increment of the objective value, we first relax the binary variable $z_u$ to be a continuous variable $\tilde{z}_u \in [0, 1]$, where $\tilde{z}_u = \min\{R_u/R_{u\min}, 1\}$. We denote $\Delta R_u$ as the increment of data rate when user $u$ is allocated additional resources, and further denote $\Delta R_{u\min} = R_{u\min} - R_u$. Then, we have

$$\Delta \tilde{z}_u = \frac{\min\{\Delta R_u, \Delta R_{u\min}\}}{R_{u\min}}. \quad (4.10)$$

We further define $\tilde{R}_{u,j,\tau} = B \log_2(1 + (P_{RRH}g_{u,j,\tau})/(\sigma^2 + \xi))$ as the unit data rate from RRH $j$ to user $u$ at the $\tau$th predicted location, and denote $\Delta w_{u,j,\tau}$ as the corresponding additional resources allocated from RRH $j$ to user $u$ at the $\tau$th predicted location. Then, for user $u$ associated with RRH $j$, the increment of data rate at the $\tau$th predicted location is

$$\Delta R_{u,j,\tau} = \Delta w_{u,j,\tau} B \log_2(1 + (P_{RRH}g_{u,j,\tau})/(\sigma^2 + \xi)) = \Delta w_{u,j,\tau} \tilde{R}_{u,j,\tau}. \quad (4.11)$$
We define \( \Delta f_{j,k} \triangleq \sum_{\tau \in T} \Delta f_{j,k,\tau} \) as the increment of the objective value when channel \( k \) is allocated to RRH \( j \), where \( \Delta f_{j,k,\tau} \) denotes the maximum increment of objective value when RRH \( j \) allocates \( \Delta t \) time slots for its associated users at the \( \tau \)th predicted locations. Thus, \( \Delta f_{j,k,\tau} \) is the optimal value of the following problem

\[
\begin{align*}
\text{maximize} & \quad \sum_{u \in A_{j,k,\tau}} (\alpha_u \Delta \tilde{z}_u(\Delta w_{u,j,\tau}) + \beta_u \Delta R_{u,j,\tau}(\Delta w_{u,j,\tau})) \\
\text{subject to} & \quad \sum_{u \in A_{j,k,\tau}} (w_{u,j,\tau} + \Delta w_{u,j,\tau}) \leq \frac{\Delta t}{T}
\end{align*}
\] (4.12a) (4.12b)

where \( A_{j,k,\tau} \) is the possible user association profile at RRH \( j \) at the \( \tau \)th prediction. We determine \( A_{j,k,\tau} \) by checking constraints (4.5d) and (4.5e). Specifically, we denote \( I_{\text{RRH}}(j,k) \) as an indicator function where \( I_{\text{RRH}}(j,k) = 0 \) indicates that constraint (4.6) cannot be satisfied if channel \( k \) is allocated to FBS \( j \), and \( I_{\text{RRH}}(j,k) = 1 \) otherwise. We denote \( I_{\text{user}}(u,j,\tau) \) as an indicator function to show whether constraint (4.5d) is satisfied at user \( u \) served by RRH \( j \) at the \( \tau \)th predicted location. We further define \( I_{\text{sp}} \) as an indicator function to show whether constraint (4.5e) is satisfied, and denote \( S' \) as the set of service providers which do not satisfy constraint (4.5e). Then, we have

\[
A_{j,k,\tau} = \begin{cases} 
\{ u \mid u \in \tilde{U}_{j,\tau}, I_{\text{user}}(u,j,\tau) = 1, s(u) \in S' \}, & \text{if } I_{\text{RRH}}(j,k) = 1 \text{ and } I_{\text{sp}} = 0, \\
\{ u \mid u \in \tilde{U}_{j,\tau}, I_{\text{user}}(u,j,\tau) = 1 \}, & \text{if } I_{\text{RRH}}(j,k) = 1 \text{ and } I_{\text{sp}} = 1, \\
\emptyset, & \text{otherwise},
\end{cases}
\] (4.13)

where \( s(u) \) is the service provider that user \( u \) subscribed to. The set in (4.13) implies that when constraint (4.5e) is not satisfied, we only consider allocating resources to users subscribed to the service providers which have not achieved the minimum guaranteed resources. Otherwise, all users are considered eligible for additional resources. It can be verified that \( \Delta \tilde{z}_u(\Delta w_{u,j,\tau}) \) is a piece-wise concave function with respect to \( \Delta w_{u,j,\tau} \). Thus,
problem (4.12) can be solved using gradient algorithm [98]. Note that we have relaxed the admission control variable $z_u$ when determining the increment of objective value. After the resource allocation process, we convert $\tilde{z}_u$ back to binary variable $z_u$ as

$$z_u = \begin{cases} 
1, & \text{if } \tilde{z}_u = 1, \\
0, & \text{otherwise.} 
\end{cases} \quad (4.14)$$

Note that the post-processing of $z_u$ does not affect the amount of resources each service provider obtained from the network operator. With the above definitions, the proposed IBGA algorithm is shown in Algorithm 4.1.

In Algorithm 4.1, we initialize all variables to be zero at the beginning in Step 1. From Steps 2 to 30, we allocate available channels to the RRHs. For each channel $k \in \mathcal{N}$, we first determine the possible set of users to be associated with each RRH in Step 3, which contains users that can be scheduled under this channel without violating the interference constraints. Then, we allocate channel $k$ iteratively to the RRHs. In each iteration, we first calculate the increment of objective value, $\Delta f_{j,k}$, according to (4.12). Then, we allocate channel $k$ to the RRH with the largest value of increment in Step 7. Next, we update the user association variables and resource sharing variables according to the solution to problem (4.12) in Step 9. Then, we update the possible user association profile $A_{j,k,\tau}$ from Steps 10 to 27. We remove the service provider $s$ from set $\mathcal{S}'$ which has achieved the minimum rate guarantee in Step 13, and update $A_{j,k,\tau}$ by removing corresponding users in the set $\mathcal{U}_s$ in Step 14. Once all the service providers achieved the minimum rate guarantee, we set $I_{sp} = 1$ in Step 18, and the procedures from Steps 10 to 17 will no longer be performed in the next iteration. By introducing steps from Steps 10 to 21, we aim to allocate resources to service providers which have not achieved the minimum rate guarantee with a higher priority. Finally, we check whether this channel can be allocated...
Algorithm 4.1 Increment-based greedy allocation (IGBA) algorithm.

1: Initialize variables \( z_s, a_s, w_s, \forall s \in S, C, f, I_{sp} \) to be all zeros and \( S' := S \).
2: \textbf{for} each channel \( k \in N \)
3: \textbf{for} each channel \( k \in N \)
4: Initialize set \( B^k := B \).
5: \textbf{while} \((B^k \neq \emptyset)\)
6: Solve problem (4.12) for \( j \in B^k \) and calculate the value of \( \Delta f_{j,k} \).
7: Find the RRH with the largest increment of objective value: \( q = \arg \max_{j \in B^k} \Delta f_{j,k} \).
8: Set \( C_{q,k} := 1 \) and update \( a_{u,q,\tau}, w_{u,q,\tau}, \forall u \in A_{q,k,\tau}, \tau \in T \) according to the solution in Step 6.
9: \textbf{if} \( I_{sp} = 0 \)
10: \textbf{for} \( s \in S' \)
11: \textbf{if} \( \sum_{u \in U_s,s \in S'} R_u \geq R_{s_{\min}} \)
12: \( S' := S' \setminus \{s\} \).
13: \textbf{endif}
14: \textbf{endfor}
15: \textbf{if} \( S' = \emptyset \)
16: Set \( I_{sp} := 1. \)
17: Update \( A_{j,k,\tau} \) according to (4.13).
18: \textbf{endif}
19: \textbf{endwhile}
20: \textbf{for} RRH \( j \in B^k \)
21: \textbf{if} \( I_{RRH}(j,k) = 0 \) or \( A_{j,k,\tau} = \emptyset, \forall \tau \in T \)
22: Set \( B^k := B^k \setminus \{j\} \).
23: \textbf{else}
24: Update \( A_{j,k,\tau} \) according to (4.13).
25: \textbf{endif}
26: \textbf{endfor}
27: \textbf{endfor}
28: \textbf{endwhile}
29: \textbf{endfor}
30: \textbf{endfor}
31: Determine the admission control variable \( z_u, u \in U_s, s \in S \) according to (4.14).

The IGBA algorithm aims to maximize the overall system utility by incrementally allocating resources to RRHs. It starts by initializing variables and sets, then iteratively assigns channels to RRHs based on an incremental criteria. If a channel cannot be assigned due to interference constraints, the algorithm updates the set of possible users and repeats the process. The algorithm ensures fairness and efficiency in resource allocation, prioritizing users with higher utility gains.
decision according to (4.14) in Step 31.

It can be seen that by using Algorithm 4.1 we can obtain a solution to problem (4.9) within a small number of iterations, i.e., $N|\mathcal{B}|$ iterations in total with respect to channel allocation. Therefore, it is more efficient compared to the standard techniques for large-scale networks. However, since we use a greedy algorithm when allocating each channel, and apply relaxation to the admission control variables, the solution obtained using Algorithm 4.1 is suboptimal. We evaluate the performance of Algorithm 4.1 in Section 4.7. By solving problem (4.9), we can obtain the resources allocated to each service provider $s \in \mathcal{S}$ at each RRH $j \in \mathcal{B}$ as $W_{j,s} = \sum_{u \in \mathcal{U}_s, \tau \in \mathcal{T}} w_{u,j,\tau}$. Then, the network operator can create the vRAN for each service provider accordingly.

### 4.5 Multi-Timescale Resource Sharing Mechanism

In the previous section, we have designed an efficient resource sharing algorithm to assist the virtualization process at the network operator, where the decisions remain unchanged for $T$ time slots. A typical challenge in designing practical virtualization mechanisms is to adapt to changes of the network status, such as the traffic variation and user mobility. Intuitively, when the users’ locations and their traffic demand do not vary, the network operator does not need to update the vRAN for each service provider. In this scenario, we can select a large value of $T$ to reduce computation and communication cost. On the contrary, when the network status changes frequently, the amount of resources required at each service provider may vary, which requires update of the vRAN frequently in a small time scale, i.e., we need to choose a small value of $T$, which results in high computation cost.

In this section, we propose a multi-timescale resource sharing mechanism as shown in Fig. 4.2 to address this issue. This mechanism consists of a global resource allocation
Figure 4.2: (a) Multi-timescale resource sharing framework: Global resource allocation performed every $T_G$ time slots, and local resource allocation performed every $T_L$ time slots. (b) Mobility prediction: predicting the locations of mobile users every $\Delta t$ time slots during the next $T_L$ time slots.

process which is performed every $T_G$ time slots, and a number of local resource allocation processes performed every $T_L = T$ time slots (where $T_G = n_L T_L$ and $n_L$ is a positive integer) between two consecutive global resource allocation. The allocated resource for each service provider remains unchanged during $T_L$ time slots. In the global resource allocation process, all users in the system are involved in the optimization, and the available resources include all the channels in set $\mathcal{N}$, as shown in problem (4.5). However, in the local resource allocation process, only users whose locations and traffic demand have changed are considered.

The available resources for the local allocation process only include the remaining resources (e.g., channels which have not been utilized) in the system. Without loss of generality, we consider global resource allocation is performed at time slot $t_0$, and the local resource allocation is performed at time slot $t_1 = t_0 + T$. All through this section, we use superscript to denote the time period when the decision is made. The remaining
resources in the system can be classified into three types. The first type is the set of channels that can further be assigned to RRHs without violating the interference constraints of the existing users, which is denoted as $\mathcal{N}_{t_1}$. The second type of remaining resources is the amount of channel resources that are assigned to some RRHs but not fully utilized by the service providers. We denote this type of remaining resources at RRH $j \in \mathcal{B}$ for the $\tau$th prediction period as $r_{j,\tau}$. The last type of remaining resources corresponds to the amount of resources released from the service providers to the network operator. This may happen when some of the users subscribed to a service provider have finished their transmission or lowered their traffic demand, and the service provider has more than enough resources to satisfy the maximum traffic demand of all its subscribed users. In this case, the service provider can decide to release the additional amount of resources back to the network operator in order to reduce expenses. We denote the amount of resources released from service provider $s$ for RRH $j$ for the $\tau$th prediction period as $\tilde{W}_{s,j,\tau}$. The optimization variables at time slot $t_1$ include $C_{j,k}^{t_1}$, $\forall k \in \mathcal{N}_{t_1}, j \in \mathcal{B}$, $z_{u}^{t_1}$, $w_{u,j,\tau}^{t_1}$, $a_{u,j,\tau}^{t_1}$, $\forall j \in \mathcal{B}_{u}^{t_1}, \tau \in \mathcal{T}, u \in \mathcal{U}_{u}^{t_1}$, $s \in \mathcal{S}$, which are defined similarly to those in the global optimization. The optimization problem for local resource allocation is similar to the global optimization problem except that the superscript $t$ is replaced by $t_1$ and constraint (4.5c) is changed to

$$\sum_{u \in \mathcal{U}_{u}} w_{u,j,\tau}^{t_1} \leq \frac{1}{n_p} \sum_{k \in \mathcal{N}_{t_1}} C_{j,k}^{t_1} + r_{j,\tau}^{t_1} + \sum_{s \in \mathcal{S}} \tilde{W}_{s,j,\tau}^{t_1}, \quad \forall \tau \in \mathcal{T}, j \in \mathcal{B}. \quad (4.15)$$

Therefore, the local optimization problem at time slot $t_1$ can be formulated as

$$\begin{align*}
\text{maximize} & \quad f^{t_1} \\
\text{subject to} & \quad (4.5b), (4.5e)-(4.5j), (4.6), \text{with superscript } t_1, \text{ and }(4.15). \quad (4.16a)\end{align*}$$

$$\begin{align*}
\text{maximize} & \quad f^{t_1} \\
\text{subject to} & \quad (4.5b), (4.5e)-(4.5j), (4.6), \text{with superscript } t_1, \text{ and}(4.15). \quad (4.16b)\end{align*}$$
Problem (4.16) can also be solved by applying Algorithm 4.1 with proper adjustment. Specifically, when initializing values of the decision variables and calculating the utility increment of the RRHs, only the amount of remaining resources and users in $\mathcal{U}_s^{t_1}$ are considered. After obtaining the solution to problem (4.16), the network operator can update the resources allocated to each service provider $s \in \mathcal{S}$ at RRH $j \in \mathcal{B}$ at time slot $t_1$ as

$$W_{s,j}^{t_1} = W_{s,j}^{t_0} - \sum_{\tau \in \mathcal{T}} \tilde{W}_{s,j,\tau}^{t_1} + \sum_{\tau \in \mathcal{T}, u \in \mathcal{U}_s^{t_1}} w_{u,j,\tau}^{t_1}.$$  (4.17)

Based on the previous discussion, the procedures of the proposed dynamic resource sharing mechanism during time slot $[t_0, t_0 + T_G)$ are shown in Algorithm 4.2. In Algorithm 4.2, the parameters $n_L = T_G/T$ and $n_p = T/\Delta t$ are predetermined integers. Steps 3 to 7 in Algorithm 4.2 correspond to the global resource allocation process, where the network operator creates a vRAN for each service provider based on its reservation requests. Step 10 is to update the locations of users in the system, which can be achieved via location monitoring techniques such as the global positioning system (GPS). In Steps 13 to 18, the network operator performs local resource update for service providers by solving problem (4.16) every $T$ time slots. In Steps 4 and 16, the network operator needs to predict the next $n_p - 1$ locations for each mobile user. Such prediction can be achieved by applying any existing mobility prediction algorithm such as the order-2 Markov predictor [97]. Note that when there is no remaining resources in the system or requests from the service providers, the network operator does not perform local optimization.

### 4.6 Further Extension

In this section, we discuss possible extensions of the proposed multi-timescale resource sharing mechanism to address some related issues of resource sharing.
Algorithm 4.2 Multi-timescale dynamic resource allocation algorithm during \([t_0, t_0 + T_G]\)

1: for \(t := t_0\) to \(T_G\)
2: \(\text{if}\ t = t_0\)
3: Collect information from service providers, including users’ locations and their QoS requirements.
4: Predict the locations of mobile users for time slot \(t + \tau \Delta t\), \(\forall \tau = 1, 2, \ldots, n_p - 1\)
5: Find resource allocation decisions by solving problem (4.9) using Algorithm 4.1.
6: Calculate the resources allocated to each service provider \(s \in \mathcal{S}\),
   \[W_{t,j}^s = \sum_{u \in \mathcal{U}_s \tau \in T} w_{u,j,\tau}^t, \forall j \in \mathcal{B}.\]
7: Create a vRAN for each service provider \(s \in \mathcal{S}\).
8: \(\text{endif}\)
9: \(\text{if}\ t = t_0 + \tau \Delta t, \forall \tau = 1, 2, \ldots \)
10: Update the location of all users in the system.
11: \(\text{endif}\)
12: \(\text{if}\ t = t_0 + mT, \forall m = 1, 2, \ldots, n_L - 1\)
13: Find the remaining channels that can further be allocated, \(N_t\), using exhaustive search.
14: Set \(r_{j,\tau}^t = 1/n_p \sum_{k \in N_t - T} C_{j,k}^t - \sum_{s \in \mathcal{S}} \sum_{u \in \mathcal{U}_s} w_{u,j,\tau}^t, \forall j \in \mathcal{B}.\)
15: Find the set of users for resource allocation, \(\mathcal{U}_s^t, \forall s \in \mathcal{S}\).
16: Find resource allocation decisions by solving problem (4.16).
17: Update the vRAN resources for each service provider \(s \in \mathcal{S}\) according to (4.17).
18: \(\text{endif}\)
19: \(\text{endfor}\)

4.6.1 Dynamic Resource Sharing for Uplink

Although the dynamic resource sharing mechanism proposed in Section 4.5 is based on downlink communication, it can also be applied to the uplink with proper adjustment. In the uplink, users transmit data to the RRHs, and each RRH may experience a different level of interference. Similar to the downlink scenario, to provide service isolation among different service providers, we restrict that the aggregate interference experienced at each RRH \(j \in \mathcal{B}\) is no larger than a threshold \(\xi'\). We also define the maximum transmission power allowed at user \(u \in \mathcal{U}_s\) as \(P_{u,\text{max}}\), which depends on the device and applications that the user is using. For simplicity, we reuse the symbols for other variables defined in the
downlink scenario. Then, the interference constraints at the RRHs are represented as

\[
C_{j,k} \sum_{l \in B \setminus \{j\}} C_{l,k} \max_{u \in \tilde{U}_l, \tau \in T} \{a_{u,l,\tau} P_{u\max} g_{u,j,\tau}\} \leq \xi', \quad \forall k \in \mathcal{N}, j \in \mathcal{B}. \tag{4.18}
\]

It can be verified that (4.18) is a non-convex constraint. To linearize (4.18), we introduce auxiliary variables \(x_{l,j}\) and \(y_{l,j,k}\) for all \(j, l \in \mathcal{B}, k \in \mathcal{N}\), where

\[
x_{l,j} = \max_{u \in \tilde{U}_l, \tau \in T} \{a_{u,l,\tau} P_{u\max} g_{u,j,\tau}\}, \tag{4.19}
\]

\[
y_{l,j,k} = C_{l,k} x_{l,j}. \tag{4.20}
\]

With the auxiliary variables, for any \(k \in \mathcal{N}\) and \(j \in \mathcal{B}\), constraint (4.18) can be transformed into the following constraints

\[
\sum_{l \in B \setminus \{j\}} y_{l,j,k} \leq \xi' + (1 - C_{j,k}) D', \tag{4.21a}
\]

\[
x_{l,j} \geq a_{u,l,\tau} P_{u\max} g_{u,j,\tau}, \quad \forall u \in \tilde{U}_l, \tau \in T, l \in B \setminus \{j\}, \tag{4.21b}
\]

\[
y_{l,j,k} \geq x_{l,j} - (1 - C_{l,k}) \max_{u \in \tilde{U}_l, \tau \in T} \{P_{u\max} g_{u,j,\tau}\}; \quad \forall l \in \mathcal{B} \setminus \{j\}, \tag{4.21c}
\]

\[
0 \leq y_{l,j,k} \leq x_{l,j}, \quad \forall l \in \mathcal{B} \setminus \{j\}; \tag{4.21d}
\]

\[
y_{l,j,k} \leq C_{l,k} \max_{u \in \tilde{U}_l, \tau \in T} \{P_{u\max} g_{u,j,\tau}\}, \quad \forall l \in \mathcal{B} \setminus \{j\}, \tag{4.21e}
\]

where \(D'\) is a large positive constant. We have the following theorem.

**Theorem 4.2** For any \(j \in \mathcal{B}\) and \(k \in \mathcal{N}\), constraint (4.18) is equivalent to constraint (4.21) if \(D'\) and \(\xi\) satisfy

\[
D' \geq (|\mathcal{B}| - 1) \max_{u \in \tilde{U}_l, \tau \in T} \{P_{u\max} g_{u,j,\tau}\} - \xi'. \tag{4.22}
\]
Proof First, we show that constraint (4.18) is equivalent to constraints (4.19)–(4.21a). It can be verified that when \( C_{j,k} = 1 \), constraint (4.18) is equivalent to constraint (4.21a) by substituting (4.19) and (4.20). When \( C_{j,k} = 0 \), constraint (4.18) is always satisfied. In this scenario, constraint (4.21a) becomes

\[
\sum_{l \in \mathcal{B} \setminus \{j\}} y_{l,j,k} \leq \xi' + D',
\]

which is also always satisfied when \( D' \geq (|\mathcal{B}|-1) \max_{u \in \mathcal{U}, s \in \mathcal{S}, \tau \in \mathcal{T}} \{ P_{\text{max}} G_{u,j,\tau} \} - \xi' \). Thus, constraint (4.18) is equivalent to constraints (4.19)–(4.21a). Next, it can be seen that (4.19) is equivalent to (4.21b). Finally, we show that (4.20) is equivalent to (4.21c)–(4.21e). When \( C_{l,k} = 0 \), we have \( y_{l,j,k} = 0 \) from (4.20). From (4.21d) and (4.21e), we also have \( y_{l,j,k} = 0 \), which implies that (4.20) is equivalent to (4.21d) and (4.21e) in this scenario. When \( C_{l,k} = 1 \), we have \( y_{l,j,k} = x_{l,j} \) from (4.20). Meanwhile, from (4.21c) and (4.21d), we also have \( y_{l,j,k} = x_{l,j} \). Therefore, (4.20) is equivalent to (4.21c)–(4.21e) under any value of \( C_{l,k} \). In summary, constraint (4.18) is equivalent to constraints (4.21a)–(4.21e). This completes the proof.

According to Theorem 4.2, we have transformed the non-convex constraint (4.18) into a set of linear constraints (4.21). Note that the objective function and all other constraints in the uplink resource allocation problem is the same as those in the downlink scenario. Therefore, the global resource allocation problem for the uplink can be formulated as

\[
\begin{align*}
\text{maximize} & \quad f \\
\text{subject to} & \quad (4.5b), (4.5c), (4.5e)-(4.5j) \text{ with } \xi = \xi', \ P_{\text{RRH}} = P_{\text{u max}}, \text{ and } (4.21). \tag{4.24b}
\end{align*}
\]

Similar to the downlink scenario, we define \( \tilde{I}_{\text{RRH}}(j,k) \) as an indicator function to show whether channel \( k \) can be allocated to RRH \( j \) without violating constraints in (4.21). We
further define $\tilde{I}_{\text{user}}(u, j, \tau)$ as the indicator function whether user $u$ can be associated with RRH $j$ in the $\tau$ prediction period without violating the interference constraints. Then, the global resource allocation problem can be solved using Algorithm 4.1 by replacing $\xi$, $P_{RRH}$, $I_{RRH}(j, k)$, and $I_{\text{user}}(u, j, \tau)$ with $\xi'$, $P_{u_{\text{max}}}$, $\tilde{I}_{RRH}(j, k)$, and $\tilde{I}_{\text{user}}(u, j, \tau)$, respectively. Similarly, we can obtain the local resource allocation decisions for the uplink using Algorithm 4.1 with the aforementioned changes. Therefore, the proposed resource sharing mechanism in Algorithm 4.2 can also be applied for uplink resource allocation.

4.6.2 Revenue Maximization for On-demand Service

In the proposed resource sharing mechanism, we optimize an objective function that characterizes the balance between throughput and the number of admitted users. In this subsection, we show that the objective function (4.4) can be extended to solve revenue maximization problem for on-demand resource sharing, where the service providers pay for the amount of resources they reserved for a certain period. We consider two different pricing schemes, a fixed-rate pricing scheme and a tiered pricing scheme, respectively. For the fixed-rate pricing scheme, we assume the price for reserving data rate $R_0$ is $\rho_0$. Then, the revenue maximization objective is

$$f_0 = \rho_0 \sum_{u \in U, s \in S} \frac{R_u}{R_0}. \quad (4.25)$$

It can be seen that (4.25) can be obtained by setting $\alpha_u = 0$ and $\beta_u = \rho_0/R_0, \forall u \in U$, $s \in S$ from (4.4). With the fixed-rate pricing scheme, the network operator’s only concern is the throughput of the system, which may result in unbalanced resource allocation among the service providers and users, e.g. users with better channel condition obtain more resources, while users with poor channel condition may not be admitted for service.
We address this issue by adopting a tiered pricing scheme. In this scheme, the price $\rho_u$ for reserving data rate $R_u$ for each user is a piece-wise function

$$
\rho_u = \begin{cases} 
0, & \text{if } R_u < R_{u\min}, \\
\frac{\rho_1 R_{u\min}}{R_0} + \frac{\rho_2 (R_u - R_{u\min})}{R_0}, & \text{if } R_{u\min} \leq R_u \leq R_{u\max}, \\
\frac{\rho_1 R_{u\min}}{R_0} + \frac{\rho_2 (R_{u\max} - R_{u\min})}{R_0}, & \text{otherwise},
\end{cases}
$$

(4.26)

where $\rho_1, \rho_2$ are constant unit prices that satisfy $\rho_1 > \rho_2$. (4.26) implies that payment is only made when the reserved data rate satisfies the user’s QoS requirement. When the data rate is greater than the upper bound $R_{u\max}$, no more payment is made for additional data rate reserved for this user. With this pricing scheme, the revenue maximization objective becomes

$$
f_1 = \sum_{u \in U_s, s \in S} \left( \frac{\rho_1 z_u R_{u\min}}{R_0} + \frac{\rho_2 z_u (R_u - R_{u\min})}{R_0} \right).
$$

(4.27)

Note that we have constraint (4.5b) to restrict the data rate for an admitted user to be within the $[R_{u\min}, R_{u\max}]$. With constraint (4.5b), (4.27) is equivalent to

$$
f_1 = \sum_{u \in U_s, s \in S} \left( \frac{(\rho_1 - \rho_2) R_{u\min}}{R_0} z_u + \frac{\rho_2 R_u}{R_0} \right),
$$

(4.28)

where we omit the $z_u$ in the second term within the bracket due to (4.5b). It can be seen that (4.28) can be obtained from (4.4) by selecting $\alpha_u = (\rho_1 - \rho_2) R_{u\min}/R_0$ and $\beta_u = \rho_2/R_0, \forall u \in U_s, s \in S$. Therefore, by adjusting the weighting factors, we can achieve different revenue optimization objectives.
Chapter 4. Resource Sharing for Cloud-based Radio Access Networks

4.7 Performance Evaluation

We evaluate the performance of the proposed dynamic resource sharing mechanism using simulation. We consider three service providers sharing a C-RAN to serve their users in a residential area. The system consists of 16 cells, where 16 RRHs are placed in a $4 \times 4$ grid. The distance between two adjacent RRHs is 30 m. There are 25 available channels, each with a bandwidth of 180 kHz. The wireless channel model follows [88]. The path loss exponent between a user and the RRH is 4. The RRH’s transmission power $P_{RRH} = 250$ mW. The noise power is $-120$ dBm. Each time slot is 100 ms. Unless specified, the users’ QoS requirements do not change during the simulation, and we set $\alpha_u = 1$, $\beta_u = 1/R_{u, \text{max}}$, $R_{s, \text{ref}} = 25$ Mbps, $\forall u \in \mathcal{U}_s$, $s \in \mathcal{S}$, $R_{\text{ref}} = 800$ kbps, $\xi = -75$ dBm, $T = 100$, $n_L = 10$, $n_p = 5$, $\Delta t = 20$ and $D = 1$. The results in this section are obtained by averaging the outcome of 50 simulation runs with different user topologies.

Figure 4.3: System throughput versus number of users per service provider.
Table 4.1: Average running time versus number of users per service provider.

<table>
<thead>
<tr>
<th>Number of users</th>
<th>Avg. running time of IBGA</th>
<th>Avg. running time of branch and bound</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>0.21 s</td>
<td>0.28 s</td>
</tr>
<tr>
<td>3</td>
<td>0.29 s</td>
<td>0.43 s</td>
</tr>
<tr>
<td>4</td>
<td>0.47 s</td>
<td>0.53 s</td>
</tr>
<tr>
<td>5</td>
<td>0.55 s</td>
<td>0.70 s</td>
</tr>
<tr>
<td>6</td>
<td>0.84 s</td>
<td>2.18 s</td>
</tr>
<tr>
<td>7</td>
<td>0.96 s</td>
<td>3.90 s</td>
</tr>
<tr>
<td>8</td>
<td>1.24 s</td>
<td>3.63×10³ s</td>
</tr>
<tr>
<td>9</td>
<td>1.43 s</td>
<td>4.06×10³ s</td>
</tr>
<tr>
<td>10</td>
<td>1.67 s</td>
<td>7.30×10³ s</td>
</tr>
</tbody>
</table>

4.7.1 Efficiency of the Proposed Algorithm

We first evaluate the efficiency of the proposed IBGA algorithm and compare its results with a standard branch and bound solution obtained using the MATLAB MILP solver. Since the complexity of the branch and bound algorithm increases significantly with respect to the size of the network, in this evaluation, we consider part of the simulation model, a $3 \times 3$ grid with 9 RRHs with 9 available channels. We vary the number of users that subscribed to each service provider from 5 to 10. The users are stationary and randomly distributed within the $100 \times 100$ m² area that covers the $3 \times 3$ grid. The QoS requirement of each user is set to be $\{800, 1000\}$ kbps. We show the average system throughput and the running time for both algorithms in Fig. 4.3 and Table 4.1, respectively. It is shown that the system throughput of both algorithms are almost the same when the number of users per service provider is less than 6. When the number of users is greater than 6, the proposed IBGA algorithm achieves no less than 92% of the performance as the branch and bound algorithm. However, the running time of the proposed IBGA algorithm is significantly lower than the MILP solver, when the number of users is greater than 7, which can be seen in Table 4.1.
4.7.2 Effect of Weighting Factors

Next, we evaluate the performance of the proposed algorithm considering the entire network (16 cells, 16 channels) with respect to different weighting factors $\alpha_u$ and $\beta_u$. In Fig. 4.4, we show the throughput of the proposed algorithm, where we vary the total number of users in the system from 72 to 156. The proposed algorithm achieves maximum throughput when $\alpha_u = 0$ and $\beta_u = 1$ and the maximum number of admitted users when $\alpha_u = 1$ and $\beta_u = 0$. When $\alpha_u = 1$ and $\beta_u = 1/R_{u\text{ max}}$, the proposed algorithm achieves the maximum number of admitted users most of the time and a throughput that is close to the maximum throughput. This justifies that by adjusting the weighting factors, we can achieve different optimization objectives during the resource allocation.

4.7.3 Effect of Interference Threshold $\xi$

Next, we evaluate the performance of the proposed mechanism with different values of interference threshold $\xi$. We fixed the number of users for each service provider as 30. Fig.
Figure 4.5: System throughput versus different interference threshold $\xi$.

4.5 shows that the system throughput first increases with $\xi$ and then decreases when $\xi$ is larger than $-75$ dBm. The reason is as follows. When $\xi$ is small, few channels can be reused among different RRHs since the interference constraints cannot be satisfied. Some users may not be admitted due to the limited amount of resources. As $\xi$ increases, channel reuse among RRHs becomes possible and more users can be admitted, which increases the system throughput. However, from (4.1), the transmission rate from a service provider to a user becomes smaller as $\xi$ increases. Thus, although the number of users admitted into the system becomes larger, the total throughput of the system may decrease when $\xi$ exceeds a certain value. In practice, $\xi$ can be selected based on simulation results under different network settings.

4.7.4 Effect of Dynamic Resource Guarantee

We compare the proposed mechanism (with IGBA algorithm) with a mechanism where each service provider are guaranteed a fixed aggregate data rate of 40 Mbps. We set
$R_{\text{ref}} = 40$ Mbps. We restrict users subscribed to service provider 3 be located in the boundary area (within 30 m to the area boundary). Fig. 4.6 shows the system throughput increases as the number of users per service provider increases. It can be seen that the two mechanisms achieve similar performance when the number of users per service provider is under 50. This is because in these scenarios, the resources in the system are sufficient to satisfy the QoS requirements of almost all users, and the total minimum rate demand for each service provider is no larger than the minimum rate guarantee (40 Mbps). Thus, using fixed or dynamic rate guarantee does not affect the system throughput. However, as the number of users increases further, the proposed mechanism with dynamic rate guarantee achieves higher throughput. The reason is that as the number of users increases, the resources become stringent and may not be sufficient to satisfy users’ QoS, especially for those subscribed to service provider 3. The proposed mechanism reduces the minimum resource guarantee for service provider 3 accordingly to save some resources for other users who are close to the RRHs. On the contrary, using fixed guarantee requires much more
resources for service provider 3, which results in inefficient utilization and reduces the system throughput.

4.7.5 Comparison with Proportional Spectrum Sharing Mechanism

Next, we compare the proposed mechanism (with IGBA algorithm) with the proportional spectrum sharing mechanism. In the proportional spectrum sharing mechanism, each RRH is allocated one channel, and this channel is shared by three service providers proportionally according to their traffic demand. The traffic demand is estimated assuming users are associated with their closest RRHs. The allocated resources remains unchanged until the next allocation process is performed. The simulation setting is the same as that in Fig. 4.6. Fig. 4.7 shows the system throughput with respect to different numbers of users per service provider. It can be seen that the proposed resource sharing mechanism achieves higher system throughput than the proportional resource sharing mechanism. This is because the
4.7.6 Effect of Traffic Variation

We let each service provider serve 30 stationary users. Each user has a QoS requirement of \{800, 1000\} kbps at the beginning. Then, every 100 time slots from time slot 100 to time slot 500, we increase the QoS requirement of service provider 1’s users by 45 kbps one at a time. Next, from time slot 600 to 1000, we decrease the QoS requirement of service provider 2’s users by 45 kbps one at a time every 100 time slots. We compare the proposed multi-timescale mechanism with a single-timescale mechanism. The single-timescale mechanism performs global optimization every 400 time slots, while the multi-timescale mechanism also performs local optimization every 100 time slots. Fig. 4.8 shows the achievable throughput over each time period with respect to the traffic variation, which is calculated according to the allocated resources. It can be seen that as the resource demand from the users varies,
the system throughput of the proposed mechanism changes accordingly every 100 time slots. This is because the proposed mechanism has a local resource allocation procedure which updates the resources allocated to each service provider every $T = 100$ time slots. However, the system throughput of the single-timescale mechanism is updated every 400 time slots due to the global resource allocation, and users’ varying resource demand between two global resource allocation processes may not be satisfied. This demonstrates that the proposed multi-timescale mechanism with local resource update can adapt to frequent traffic variation and can provide on-demand services.

### 4.7.7 Effect of User Mobility

We consider each service provider serves 15 stationary users and 15 mobile users each with QoS requirement \{600, 800\} kbps. We adopt the 2D Gauss-Markov movement model [99], where the velocity of a mobile user is correlated in time. User $u$’s velocity in each dimension $v^t_u$ at time $t$ is given by $v^t_u = \gamma_v v^{t-1}_u + (1 - \gamma_v) \mu_v + \sqrt{1 - \gamma_v^2} x^{t-1}$, where $\gamma_v \in [0, 1]$ is the velocity memory factor, $\mu_v$ is the asymptotic mean of $v^t_u$ and $x$ is an independent and stationary Gaussian random variable with zero mean and standard deviation $\sigma_v$. We set $\gamma_v = 0.9$, $\mu_v = 1$ m/s, and $\sigma_v = 1$. The initial speed of each user is selected as 0.25 m/s at a random direction and is updated every four seconds. For location prediction, we implement an order-2 Markov predictor [97], which predicts the next movement in one direction based on the most recent two movements in the same direction. Each movement is measured by a speed change in the horizontal direction and the vertical direction, respectively, with step size selected from $[-2.5, -2.25, \ldots, 2.25, 2.5]$ m/s. The parameter of the Markov predictor is obtained via simulation for $10^5$ time slots. We predict four future positions for each user during the optimization process. Fig. 4.9 shows the system throughput with respect to different average speeds of mobile users. The system throughput decreases for
both mechanisms with or without mobility prediction. This is because as the users move faster, the amount of resources required at each service provider from different RRHs varies more quickly. However, the amount of resources allocated to each service provider remain unchanged for a certain period, which may not be sufficient to satisfy the QoS requirements of all mobile users. Thus, the system throughput decreases since some mobile users are not admitted for services. Nevertheless, with the mobility prediction, the proposed mechanism achieves around 10% throughput improvement as the users move faster than 3.5 m/s, as shown in Fig. 4.9.

We also evaluate the effect of prediction error on the propose mechanism, where we manually add prediction error to the actual locations of users during each prediction period. Fig. 4.10 shows the system throughput with respect to different prediction errors with users’ average speed of 2 m/s. It can be seen that as the error increases, the system throughput decreases. However, the decrease is within 1% as long as the prediction error is within 6 m.
4.8 Summary

In this chapter, a multi-timescale dynamic resource sharing mechanism was proposed. In this mechanism, the network operator performed a global resource allocation at a relatively large time scale, and performed local resource allocations based on the changes of network status such as traffic variation and user mobility. A threshold-based policy was adopted to limit the aggregate interference experienced at each user when optimizing the resource allocation decisions. This threshold policy provided service isolation among the service providers. The proposed mechanism defined a dynamic resource guarantee metric to improve the efficiency of resource utilization. It also employed a mobility prediction approach to facilitate the estimation of traffic demand. An efficient algorithm was proposed to find the desired resource allocation decision for the network operator and service providers. This chapter also discussed possible extensions of the proposed mechanism for uplink communication and revenue maximization. Through simulations, it was shown that the proposed mechanism achieved service isolation and efficient resource
sharing among service providers. It adapted to traffic variation, and achieved robust performance under user mobility. The proposed mechanism achieved a higher system throughput than the proportional spectrum sharing mechanism.
Chapter 5

Conclusions and Future Work

In this chapter, we summarize the results and highlight the contributions of this thesis. We also suggest several topics for future work.

5.1 Research Contributions

- Chapter 2 proposed a downlink scheduling mechanism with transmission strategy selection for multi-cell MIMO networks. This mechanism dynamically selects a user and the corresponding transmission strategy for each base station in each time slot to maximize the overall system utility while stabilizing all data queues. It was shown that this mechanism achieved superior performance than the scheduling mechanisms with a single transmission strategy. The proposed mechanism can be implemented at a central scheduler which performs scheduling for the entire network. It can also be implemented distributively at each base station with coordination, i.e., each base station only needs to share the value of one variable. Since the proposed scheduling mechanism optimizes long term system utility while guaranteeing the stability, it is suitable for delay-tolerant applications, such as file downloading, where system stability and long term performance are more significant.

- Chapter 3 studied uplink resource allocation problem for a two-tier macro-femto network. A five-stage network configuration mechanism was proposed which consists of access control, channel allocation, and power management processes.
The contributions for this mechanism can be explained from two aspects. First, this mechanism provides incentive for hybrid access operation at FBSs using a resource allocation approach, which does not require changing the business model of the system. Second, this mechanism coordinates different decision makers and optimizes their decisions with different objectives accordingly. It was shown that the proposed mechanism achieved a higher system utility and throughput compared to the mechanisms with the closed access scheme or a topology-based access control with orthogonal channel allocation scheme. This mechanism can provide efficient network configuration and can be applied for networks deployed in urban residential areas or commercial districts, where hybrid access is desirable to improve the system performance.

- Chapter 4 proposed a multi-timescale resource sharing mechanism for C-RANs to support multiple service providers who lease radio resources from a network operator. This mechanism consists of a global resource allocation process and several local resource allocation processes to deal with the variation of the network status, which is more efficient than performing network-wide optimization only. In this mechanism, the network operator jointly determines the resource to be allocated to each service provider as well as the user admission and association decisions. There are three major advantages of this mechanism. First, the proposed mechanism guarantees service isolation among different service providers while utilizing the spectrum resources efficiently. This is achieved through optimization with an interference threshold policy. Second, the proposed mechanism is robust to traffic variation and user mobility due to use of multi-timescale framework and mobility prediction during the optimization. Finally, this mechanism can provide on-demand resource sharing among service providers and is able to deal with traffic
overload using admission control. It was shown that the proposed mechanism achieved efficient resource utilization and service isolation among the service providers under various network situations with different traffic loads.

5.2 Suggestions for Future Work

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

1. **Uplink random access for multi-cell MIMO networks**: In Chapter 2, a downlink scheduling mechanism with transmission strategy selection was proposed for multi-cell MIMO networks. However, the uplink random access problem was not studied. In multi-cell networks, designing an uplink random access mechanism is challenging due to inter-cell interference. Packet collision is not only caused by contention among users within a cell, but also caused by interference from users in adjacent cells. The conventional random access mechanisms, such as the slotted aloha mechanism and carrier sense multiple access (CSMA) mechanism, resolve collision by adjusting transmission probability or backoff window, while interference management techniques are not considered. It is interesting to analyze the impact of interference on the performance of conventional random access schemes. Interference management techniques such as interference alignment and cancellation can be applied to design a new random access mechanism.

2. **Pricing scheme to motivate hybrid access**: In Chapter 3, a resource allocation approach is employed to motivate hybrid access mode at the FBSs in a two-tier macro-femto networks. This approach is suitable for the scenario where using hybrid access mode benefits both the FUs and MUs. However, in certain scenarios, using hybrid access may not bring higher data rates for FUs due to limited resources.
Chapter 5. Conclusions and Future Work

To further motivate the use of hybrid access mode, a suitable pricing scheme may be needed. Such a pricing scheme may also affect the resource allocation process. It would be interesting to design and analyze a novel pricing scheme for two-tier heterogeneous networks.

3. The effect of joint transmission on resource sharing in C-RAN: In Chapter 4, a resource sharing mechanism was proposed for C-RANs. In this mechanism, it is assumed that a user can only be served by one RRH at a time. It has been shown that using joint transmission from multiple RRH can significantly improve the signal quality, especially for mobile users. By employing joint transmission, the resources at a set of RRHs should be reserved simultaneously for a particular user, which may reduce diversity gain. The effect of using joint transmission in designing resource sharing mechanism is a possible direction to extend the work in Chapter 4.
Bibliography


Appendix A

Proof of Proposition 2.1

First, with the auxiliary variables $\gamma[t] = (\gamma_{ik}[t], k \in \mathcal{K}_i, i \in \mathcal{I})$, we introduce the following problem:

\[
\begin{align*}
\text{maximize} & \quad \phi(\gamma) \\
\text{subject to} & \quad \sum_{k \in \mathcal{K}_i} l_k[t] = 1, \quad \forall \ i \in \mathcal{I}, \ t \in \{1, 2, \ldots\}, \\
& \quad R_{ik} \geq \overline{A}_{ik}, \quad \forall \ k \in \mathcal{K}_i, \ i \in \mathcal{I}, \\
& \quad \overline{\gamma}_{ik} \leq R_{ik}, \quad \forall \ k \in \mathcal{K}_i, \ i \in \mathcal{I}, \\
& \quad 0 \leq \gamma_{ik} \leq \gamma_{\max}, \quad \forall \ k \in \mathcal{K}_i, \ i \in \mathcal{I}, \\
\end{align*}
\]

(A.1)

where

\[
\phi(\gamma) = \lim_{t \to \infty} \frac{1}{t} \sum_{\tau=0}^{t-1} \phi(\gamma),
\]

(A.2)

and

\[
\overline{\gamma}_{ik} = \lim_{t \to \infty} \frac{1}{t} \sum_{\tau=0}^{t-1} \gamma_{ik}[\tau].
\]

(A.3)

According to [23, Chapter 5], it can be proved that designing a policy to solve problem (A.1) ensures all the desired constraints of the original problem are satisfied while providing a utility that is at least as good as $\phi(\overline{R}^*(\gamma_{\max}))$, where $\overline{R}^*(\gamma_{\max})$ is the solution to the original problem (2.14) with the additional constraint $0 \leq \overline{R}_{ik} \leq \gamma_{\max}, \ \forall \ k \in \mathcal{K}_i, \ i \in \mathcal{I}$.

We denote $\Theta[t] = [Q[t], W[t]]$, and define a Lyapunov function:

\[
L(\Theta[t]) = \frac{1}{2} \left( \sum_{i \in \mathcal{I}} \sum_{k \in \mathcal{K}_i} Q_{ik}[t]^2 + \sum_{i \in \mathcal{I}} \sum_{k \in \mathcal{K}_i} W_{ik}[t]^2 \right).
\]

(A.4)
Appendix A. Proof of Proposition 2.1

Then, we have

\[
L(\Theta[t+1]) - L(\Theta[t])
= \frac{1}{2} \sum_{i \in I} \sum_{k \in K_i} (Q_{ik}[t+1]^2 - Q_{ik}[t]^2) + \frac{1}{2} \sum_{i \in I} \sum_{k \in K_i} (W_{ik}[t+1]^2 - W_{ik}[t]^2)
= \frac{1}{2} \sum_{i \in I} \sum_{k \in K_i} ((\max[Q_{ik}[t] - R_{ik}[t], 0] + A_{ik}[t])^2 - Q_{ik}[t]^2)
+ \frac{1}{2} \sum_{i \in I} \sum_{k \in K_i} ((\max[W_{ik}[t] - R_{ik}[t], 0] + \gamma_{ik}[t])^2 - W_{ik}[t]^2)
\leq \sum_{i \in I} \sum_{k \in K_i} \frac{1}{2} (A_{ik}[t]^2 + \gamma_{ik}[t]^2 + 2R_{ik}[t]^2)
+ \sum_{i \in I} \sum_{k \in K_i} Q_{ik}[t] (A_{ik}[t] - R_{ik}[t]) + \sum_{i \in I} \sum_{k \in K_i} W_{ik}[t] (\gamma_{ik}[t] - R_{ik}[t]). \tag{A.5}
\]

We define a Lyapunov drift as

\[
\Delta(\Theta[t]) \triangleq \mathbb{E}\{L(\Theta[t+1]) - L(\Theta[t]) \mid \Theta[t]\}.
\]

It can be shown that

\[
\Delta(\Theta[t])
\leq \mathbb{E}\left\{\sum_{i \in I} \sum_{k \in K_i} \frac{1}{2} (A_{ik}[t]^2 + \gamma_{ik}[t]^2 + 2R_{ik}[t]^2) \mid \Theta[t]\right\}
+ \mathbb{E}\left\{\sum_{i \in I} \sum_{k \in K_i} Q_{ik}[t] (A_{ik}[t] - R_{ik}[t]) \mid \Theta[t]\right\}
+ \mathbb{E}\left\{\sum_{i \in I} \sum_{k \in K_i} W_{ik}[t] (\gamma_{ik}[t] - R_{ik}[t]) \mid \Theta[t]\right\}. \tag{A.6}
\]

Now, we define \( D \) as a finite constant that bounds the first term on the right-hand-side of the above drift inequality, so that for all \( t \), all possible \( \Theta[t] \), and all possible control actions
that can be taken, we have

$$
E \left\{ \sum_{i \in I} \sum_{k \in \mathcal{K}_i} \frac{1}{2} \left( A_{ik}[t]^2 + \gamma_{ik}[t]^2 + 2R_{ik}[t]^2 \right) \mid \Theta[t] \right\} \leq D. \quad (A.7)
$$

Assuming i.i.d. channels, from (A.6), we have

$$
\Delta(\Theta[t]) - \beta E \{ \phi(\gamma[t]) \mid \Theta[t] \} \leq D - \beta E \{ \phi(\gamma[t]) \mid \Theta[t] \} + E \left\{ \sum_{i \in I} \sum_{k \in \mathcal{K}_i} Q_{ik}[t](A_{ik}[t] - R_{ik}[t]) \mid \Theta[t] \right\} + E \left\{ \sum_{i \in I} \sum_{k \in \mathcal{K}_i} W_{ik}[t](\gamma_{ik}[t] - R_{ik}[t]) \mid \Theta[t] \right\}. \quad (A.8)
$$

The proposed stochastic optimization steps (i)-(iii) in Section 2.4.1 aim to minimize the right hand side of the above inequality given any realization of $\Theta[t]$, which leads to

$$
\Delta(\Theta[t]) - \beta E \{ \phi(\gamma[t]) \mid \Theta[t] \} \leq D - \beta \phi(\gamma^*) + E \left\{ \sum_{i \in I} \sum_{k \in \mathcal{K}_i} Q_{ik}[t](A_{ik}^*[t] - R_{ik}^*[t]) \mid \Theta[t] \right\} + E \left\{ \sum_{i \in I} \sum_{k \in \mathcal{K}_i} W_{ik}[t](\gamma_{ik}^* - R_{ik}^*[t]) \mid \Theta[t] \right\}, \quad (A.9)
$$

where $\gamma^* = (\gamma_{ik}^*, k \in \mathcal{K}_i, i \in I)$ is any vector in the feasible region, $A_{ik}^*$ is any arrival rate, and $R_{ik}^*$ is derived from any feasible scheduling policy. It has been shown in [23, Chapter 5] that if a feasible solution of the original problem exists, for any $\delta > 0$, there is a feasible scheduling policy and a vector $\gamma^*$ such that

$$
-\phi(\gamma^*) \leq -\phi(\bar{R}^*(\gamma_{\text{max}})) + \delta, \quad (A.10)
$$
Appendix A. Proof of Proposition 2.1

\[ E \{ A_{ik}^*[t] - R_{ik}^*[t] \} \leq \delta, \quad \forall \ k \in K, \ i \in I, \quad (A.11) \]

\[ E \{ (\gamma_{ik}^*[t] - R_{ik}^*[t]) \} \leq \delta, \quad \forall \ k \in K, \ i \in I. \quad (A.12) \]

Taking \( \delta \to 0 \), together with (A.9), we have

\[ \Delta(\Theta[t]) - \beta E \{ \phi(\gamma[t]) \mid \Theta[t] \} \leq D - \beta \phi(\overline{R}^*(\gamma_{\text{max}})). \quad (A.13) \]

By applying the telescoping sums in the above inequality (where we take the value of \( t \) in (A.13) from 0 to \( t - 1 \) and take the summation of both sides of all the inequalities), for all \( t > 0 \), we have

\[ \frac{1}{t} \sum_{\tau=0}^{t-1} E\{\phi(\gamma[\tau]) \mid \Theta[\tau] \} \geq \phi(\overline{R}^*(\gamma_{\text{max}})) - \frac{D}{\beta} - \frac{E[L(\Theta(0))] \beta t}{t}. \quad (A.14) \]

According to Jensen’s inequality for the concave function \( \phi(\cdot) \), we have

\[ \liminf_{t \to \infty} \phi(\overline{\gamma}) \geq \phi(\overline{R}^*(\gamma_{\text{max}})) - \frac{D}{\beta}. \quad (A.15) \]

On the other hand, rearranging (A.13) yields

\[ \Delta(\Theta[t]) \leq D + \beta \left( E \{ \phi(\gamma[t]) \mid \Theta[t] \} - \phi(\overline{R}^*(\gamma_{\text{max}})) \right). \quad (A.16) \]

According to the Lyapunov drift theorem in [23, Chapter 1], (A.16) implies that all the queues are mean rate stable, which means \( \overline{\gamma}_{ik} - \overline{R}_{ik} \leq 0 \). Using this along with the continuity and entrywise non-decreasing properties of \( \phi(\cdot) \), we have

\[ \liminf_{t \to \infty} \phi \left( \frac{1}{t} \sum_{\tau=0}^{t-1} E[R[\tau]] \right) \geq \liminf_{t \to \infty} \phi(\overline{\gamma}) \geq \phi(\overline{R}^*(\gamma_{\text{max}})) - \frac{D}{\beta}. \quad (A.17) \]
Appendix A. Proof of Proposition 2.1

Note that we assume there exists a feasible scheduling scheme $\nu$ which leads to $0 \leq \mathbb{E}\{R^{\nu}_{ik}[t]\} \leq \gamma_{\max}$, $\mathbb{E}\{A_{ik}[t] - R^{\nu}_{ik}[t]\} \leq -\epsilon$, and $\phi(\mathbb{E}\{R^{\nu}_{ik}[t]\}) = \phi_{\epsilon}$. From (A.9), we have

$$\Delta(\Theta[t]) \leq D + \beta \left( \mathbb{E}\left\{\phi(\gamma[t]) \mid \Theta[t]\right\} - \phi_{\epsilon}\right) - \epsilon \sum_{i \in I} \sum_{k \in K_{i}} Q_{ik}[t].$$  \hspace{1cm} (A.18)

By rearranging the above inequality and taking iterated expectations and telescoping sums, we obtain

$$\frac{1}{t} \sum_{\tau=0}^{t-1} \sum_{i \in I} \sum_{k \in K_{i}} \mathbb{E}\{Q_{ik}[\tau]\} \leq \frac{D + \beta(\frac{1}{t} \sum_{\tau=0}^{t-1} \mathbb{E}\{\phi(\gamma[\tau])\} - \phi_{\epsilon})}{\epsilon} - \frac{\mathbb{E}\{L(\Theta(0))\}}{\epsilon t}. \hspace{1cm} (A.19)$$

Since $\lim_{t \to \infty} \frac{1}{t} \sum_{\tau=0}^{t-1} \mathbb{E}\{\phi(\gamma[\tau])\} \leq \phi(\overline{R}(\gamma_{\max}))$, taking $\lim_{t \to \infty} \sup$ at both sides of (A.19), we have

$$\limsup_{t \to \infty} \frac{1}{t} \sum_{\tau=0}^{t-1} \sum_{i \in I} \sum_{k \in K_{i}} \mathbb{E}\{Q_{ik}[\tau]\} \leq \frac{D + \beta(\phi(\overline{R}(\gamma_{\max})) - \phi_{\epsilon})}{\epsilon}. \hspace{1cm} (A.20)$$

This completes the proof. \hfill \blacksquare
Appendix B

Proof of Theorem 3.1

According to the discussion in Section 3.4.2, we only need to prove that Algorithm 3.1 solves subproblems (3.12) and (3.13). First, it is clear that the solution to subproblem (3.12) is to allocate the minimum amount of resource to the MUs that guarantees their QoS constraints, which corresponds to Lines 1 and 2 in Algorithm 3.1.

Next, we show that Algorithm 3.1 also solves subproblem (3.13) optimally. We denote $\Delta \tilde{u}_{ij}(k) = \Delta u_{ij}(\frac{k}{T})$ as the utility increment when FU $i$ obtains its $k$th time slot from the FBS. It can be shown that $\Delta \tilde{u}_{ij}(k)$ is a non-decreasing function of $k$, which gives

$$\Delta \tilde{u}_{ij}(k) \geq \Delta \tilde{u}_{ij}(k + 1) \quad \text{if } w_{ij}^{\min T} < k < T. \quad (B.1)$$

The proof of (B.1) is shown in Appendix C. We denote set $\mathcal{K} = \{1, 2, \ldots, T\}$, and define $W_{ij}^k$ as an indicator such that $W_{ij}^k = 1$ indicates user $i$ obtains its $k$th time slot ($W_{ij}^k = 0$ otherwise). Then, subproblem (3.13) is equivalent to the following problem

$$\begin{align*}
\text{maximize} & \quad \sum_{i \in S_f^j} \sum_{k \in \mathcal{K}} W_{ij}^k \Delta \tilde{u}_{ij}(k) \\
\text{subject to} & \quad \sum_{i \in S_f^j} \sum_{k \in \mathcal{K}} W_{ij}^k \leq T \left( \sum_{k \in \mathcal{K}} x_j^k - \sum_{i \in S_m^j} w_{ij} \right), \\
& \quad W_{ij}^k = 1, \quad \forall k \leq w_{ij}^{\min T}, \forall i \in S_f^j, \\
& \quad \prod_{t=1}^{k} W_{ij}^t = W_{ij}^k, \quad \forall k \in \mathcal{K}, \forall i \in S_f^j. \quad (B.2)
\end{align*}$$
Appendix B. Proof of Theorem 3.1

The first constraint indicates the total number of time slots allocated to the FUs is no larger than \( T(\sum_{k \in N} x_j^k - \sum_{i \in S_j} w_{ij}) \). The second constraint indicates that user \( i \) should obtain at least \( w_{ij}^{\text{min}} T \) time slots from the FBS. The third constraint implies that if user \( i \) obtains its \( k \)th time slot from the FBS \( (W_{ij}^k = 1) \), it should also obtain all the previous time slots \( (W_{ij}^t = 1, \forall t \leq k) \). We define the set \( \Delta U_j = \{ \Delta \tilde{u}_{ij}(k), \forall k \in \{w_{ij}^{\text{min}} T + 1, \ldots, T\}, \forall i \in S_j^f \} \) and denote \( \Delta U_j' \) as the set of the largest \( T_R \) (defined in Algorithm 3.1) elements from \( \Delta U_j \).

With property (B.1), it can be shown that if \( \Delta \tilde{u}_{ij}(k) \in \Delta U_j' \), then \( \Delta \tilde{u}_{ij}(t) \in \Delta U_j', \forall t < k \). Therefore, the corresponding values \( W_{ij}^k = 1 \) associated with the positive elements \( \Delta \tilde{u}_{ij}(k) \) in \( \Delta U_j' \) constitute the solution to (B.2). Note that in Algorithm 3.1, for each remaining time slot after basic allocation, we allocate it to the user with the largest positive \( \Delta u_{ij}(w_{ij} + 1/T) \).

Based on the non-increasing property of \( \Delta u_{ij}(w_{ij}) = \Delta \tilde{u}_{ij}(w_{ij} T) \), Algorithm 3.1 exactly finds the positive elements among the first largest \( T_R \) elements in set \( \Delta U_j \). Therefore, Algorithm 3.1 also solves problem (B.2), which implies that it solves subproblem (3.13). This completes the proof. ■
Appendix C

Proof of Inequality (B.1)

According to the definition of $\Delta \tilde{u}_{ij}(k)$, to prove inequality (B.1), we only need to prove that $u_{ij}(w_{ij}, P_i^*(w_{ij}))$ is a concave function with respect to $w_{ij}$ in $(w_{ij}^{\min}, 1)$ (where we relax the integer constraint on $w_{ij}$). We define $\phi(w_{ij}) = \frac{du_{ij}}{dw_{ij}}$ and show that $\frac{d\phi}{dw_{ij}} < 0, \forall w_{ij} \in (w_{ij}^{\min}, 1)$. Note that according to (3.10) in Section 3.4.1, the optimal power $P_i^*$ is a function of $w_{ij}$, which may be either $P_i^{\min}$, $P_{\max}$ or $\tilde{P}_i$. We define $W_1, W_2$ and $W_3$ as the intervals of $w_{ij}$ where $P_i^*$ is $P_i^{\min}$, $P_{\max}$ and $\tilde{P}_i$, respectively. We show that for $w_{ij} \in W_i, i \in \{1, 2, 3\}$, we always have $\frac{d\phi}{dw_{ij}} < 0$.

First, if $W_1 \neq \emptyset$, we have $P_i^* = P_i^{\min}$ and $\tilde{R}_{ij} = R_i^{\min}$ for $w_{ij} \in W_1$. In this case, $u_{ij} = C - \beta w_{ij}\sigma^2 + \xi_j \left(2R_i^{\min}/(w_{ij}B) - 1\right)$. Then, we have

$$\phi(w_{ij}) = -\frac{\beta(\sigma^2 + \xi_j)}{g_{ij}} \cdot 2 \left(R_i^{\min}/w_{ij}\right) \left(1 - \frac{R_i^{\min} \ln 2}{Bw_{ij}}\right) + 1,$$

and

$$\frac{d\phi}{dw_{ij}} = -\frac{\beta(\sigma^2 + \xi_j)}{g_{ij}} \cdot 2 \left(R_i^{\min}/w_{ij}\right)^2 \cdot 2 \left(R_i^{\min}/Bw_{ij}\right) < 0.$$

Next, if $W_2 \neq \emptyset$, we have $P_i^* = P_{\max}$, and $\tilde{R}_{ij} = w_{ij}B \log_2(1 + \frac{q_{i\max}}{\sigma^2 + \xi_j})$ for $w_{ij} \in W_2$. Then, we have $u_{ij} = \Phi_i(\tilde{R}_{ij}) - \beta w_{ij}P_{\max}$ and $\phi(w_{ij}) = \frac{d\Phi_i}{d\tilde{R}_{ij}} \cdot B \log_2(1 + \frac{q_{i\max}}{\sigma^2 + \xi_j}) - \beta P_{\max}$. Since $\Phi_i(\cdot)$ is a concave function of $\tilde{R}_{ij}$, $\frac{d\Phi_i}{d\tilde{R}_{ij}} < 0$. Note that in this case, $\tilde{R}_{ij}$ is an increasing function of $w_{ij}$. Therefore, $\phi(w_{ij})$ decreases with $w_{ij}$, which gives $\frac{d\phi}{dw_{ij}} < 0$.

Finally, if $W_3 \neq \emptyset$, we have $P_i^* = \tilde{P}_i$ for $w_{ij} \in W_3$, where $\tilde{P}_i$ satisfies $\frac{d\phi}{d\tilde{R}_{ij}} \frac{d\tilde{R}_{ij}}{dP_i} - \beta w_{ij} = 0$. 

141
Then, we have
\[ \frac{d\Phi_i}{dR_{ij}} \bigg|_{P_i = \tilde{P}_i} B \ln 2 \left( \tilde{P}_i + \frac{\sigma^2 + \xi_j}{g_{ij}} \right) - \beta = 0. \]  
(C.3)

In this case, we first show that \( \tilde{P}_i \) decreases with \( w_{ij} \). Assume \( w'_{ij} \) and \( \tilde{P}'_i \) also satisfy (C.3) and \( w'_{ij} > w_{ij} \). If \( \tilde{P}'_i \geq \tilde{P}_i \), we have \( \tilde{R}'_{ij} \geq \tilde{R}_{ij} \). Since \( \frac{d\Phi_i}{dR_{ij}} \) decreases with \( \tilde{R}_{ij} \), we have
\[ \frac{d\Phi_i}{dR_{ij}} \bigg|_{P_i = \tilde{P}_i} B \ln 2 \left( \tilde{P}_i + \frac{\sigma^2 + \xi_j}{g_{ij}} \right) - \beta > \frac{d\Phi_i}{dR_{ij}} \bigg|_{P_i = \tilde{P}'_i} B \ln 2 \left( \tilde{P}'_i + \frac{\sigma^2 + \xi_j}{g_{ij}} \right) - \beta = 0, \]  
(C.4)

which contradicts (C.3). Therefore, for \( w'_{ij} > w_{ij} \), we must have \( \tilde{P}'_i < \tilde{P}_i \), which implies that \( \frac{d\tilde{P}_i}{dw_{ij}} < 0 \).

We further have
\[
\phi(w_{ij}) = \frac{d\Phi_i}{dR_{ij}} \bigg|_{P_i = \tilde{P}_i} \left( B \log_2 \left( 1 + \frac{g_{ij} \tilde{P}_i}{\sigma^2 + \xi_j} \right) \right) + w_{ij} \frac{B}{\ln 2 \left( \tilde{P}_i + (\sigma^2 + \xi_j)/g_{ij} \right)} \cdot \frac{d\tilde{P}_i}{dw_{ij}} - \beta \left( \tilde{P}_i + w_{ij} \frac{d\tilde{P}_i}{dw_{ij}} \right) = \beta \ln 2 \left( \tilde{P}_i + \frac{\sigma^2 + \xi_j}{g_{ij}} \right) \log_2 \left( 1 + \frac{g_{ij} \tilde{P}_i}{\sigma^2 + \xi_j} \right) - \beta \tilde{P}_i. \]  
(C.5)

From (C.5), we have \( \frac{d\phi}{dw_{ij}} = \frac{d\phi}{d\tilde{P}_i} \frac{d\tilde{P}_i}{dw_{ij}} = \beta \ln 2 \log_2 (1 + \frac{g_{ij} \tilde{P}_i}{\sigma^2 + \xi_j}) \frac{d\tilde{P}_i}{dw_{ij}} < 0 \) according to \( \frac{d\tilde{P}_i}{dw_{ij}} < 0 \).

In summary, we have shown that \( \frac{d\phi}{dw_{ij}} < 0 \) for \( w_{ij} \in W_1 \cup W_2 \cup W_3 \). Note that \( (w_{ij}^{\min}, 1) \subseteq W_1 \cup W_2 \cup W_3 \), and \( \phi(w_{ij}) \) is a continuous function over \( (w_{ij}^{\min}, 1) \). Therefore, we conclude that \( \frac{d\phi}{dw_{ij}} < 0, \forall w_{ij} \in (w_{ij}^{\min}, 1) \), which implies \( u_{ij}(w_{ij}, P_i^*(w_{ij})) \) is a concave function of \( w_{ij} \) in \( (w_{ij}^{\min}, 1) \). This completes the proof.