Towards Efficient 5G Wireless Networks: Cross-Layer Design and Wireless Networking

by

Mingjie Feng

A dissertation submitted to the Graduate Faculty of
Auburn University
in partial fulfillment of the
requirements for the Degree of
Doctor of Philosophy

Auburn, Alabama
August 4, 2018

Keywords: 5G Wireless Network, Massive MIMO, Millimeter Wave, Cross-Layer Design

Approved by

Shiwen Mao, Chair, Samuel Ginn Endowed Professor of Electrical and Computer Engineering
Robert Nelms, Professor and Chair of Electrical and Computer Engineering
Jitendra Tugnait, James B Davis Professor of Electrical and Computer Engineering
Xiaowen Gong, Assistant Professor of Electrical and Computer Engineering

Abstract

With the fast growing popularity of smart mobile devices and the explosion of data-intensive services, the 5th generation (5G) wireless system is expected to provide a 1000x mobile data rate in the near future. To provide high data rate services to a large number of user devices, possible approaches include aggressive spectrum reuse, highly efficient multiplexing, and increased spectrum bandwidth. To this end, several promising technologies were proposed including massive MIMO (Multiple Input Multiple Output), small cell, and mmWave (Millimeter Wave) communication.

However, the successful applications of these emerging technologies face new challenges. For example, the high channel estimation overhead of massive MIMO systems due to large number of antennas; the interference issue in small cells network due to the dense deployment of base stations (BS); the vulnerability to blockage for mmWave communication due to short wavelength. To harvest the benefits of each technology, the 5G systems are expected to be an integration of multiple technologies. However, due to the inherent limits of each technology, such integration faces new challenges which need to be addressed with proper design.

In this dissertation work, we aim to address the key challenges of 5G emerging wireless systems from the perspectives of cross-layer design and wireless networking. Through analyzing the special properties of different technologies, we propose solutions to enable efficient integration and enhance the system performance.

The first part of this dissertation investigates the problem of dynamic BS sleep control for energy efficient massive MIMO heterogeneous network (HetNet). To achieve a good balance between data rate and energy consumption, we aim to maximize the energy efficiency of a massive MIMO HetNet through dynamic BS ON-OFF switching. Such a problem is formulated as an integer programming and we proposed centralized and distributed schemes to determine the set of small cell BSs (SBS) to be turned off.

The second part of this dissertation presents a solution of interference management in a massive MIMO HetNet with non-uniform antenna placement. We apply an antenna array configuration and processing technique called nested array for interference management. The design issue is how to use the degree of freedom (DoF) to serve users as well as nullify interference such that the network performance can be optimized. We proposed effective solutions and demonstrated the improved performance of proposed schemes with simulations.

The third part of this dissertation presents a cross layer design for wireless backhaul-based massive MIMO HetNet. We consider a joint frame design, resource allocation, and user association scheme to maximize the sum rate of a massive MIMO HetNet with wireless backhaul. We formulate the problem as an integer programming and propose an iterative solution algorithm. We show that with adaptive pilot length, i.e., the number of symbols dedicated to pilots in each frame, the system performance can be enhanced compared to the schemes with fixed frame structures.

The fourth part of this dissertation investigates the problem of duplex mode selection and resource allocation for full-duplex enabled femtocell networks. We first employ a stable roommate matching algorithm to determine the pairing strategy of users and make a selection between half duplex and full duplex based on the pairing result. We then consider channel and power allocation with the objective improving the sum rate. With the proposed scheme, the interference caused by full duplex transmission is effectively controlled and the system performance is improved compared to the cases without user pairing and adaptive mode selection.

The fifth part of this dissertation presents a BS cooperation architecture for providing high-data-rate service to large number of users with dense deployment of small cells. We propose a cooperative small cell network architecture to mitigate inter-cell interference and improve the network capacity. The key components include adaptive BS deployment and configuration, dynamic resource allocation, and interference coordination. With efficient spectrum reuse and traffic-aware scheduling, the proposed architecture effectively improves the data rate performance of small cell network when serving large number of users.

The sixth part of this dissertation presents a solution of dealing with link blockage with a combination of multiple approaches. We consider a combination of device to device (D2D)

relaying and multi-beam reflection to enhance the performance of a mmWave system serving large number of users. We designed an adaptive scheme to select the set of users served by each approach. Simulation results demonstrated the performance gain achieved by adopting a combination of multiple approaches to overcome blockage.

Acknowledgments

First, I would like to express my sincere gratitude to my major advisor Prof. Shiwen Mao for his great support and helpful guidance during my Ph.D. study. With extensive knowledge and deep understanding of the research area, Prof. Mao inspired me in many ways and helped me to overcome the difficulties I met, through which I received a good training in various aspects. His high level of expertise and continuous encouragement motivated me to devote myself to research and pursue an academic career. I always feel lucky to be his student. In addition, I would like to thank Prof. Mao's wife, Prof. Yihan Li for her care and encouragement in the past five years.

Besides my advisor, I would like to thank the other committee members of my dissertation: Prof. Robert Nelms, Prof. Jitendra Tugnait, and Prof. Xiaowen Gong, and the University Reader Prof. Tao Shu, for their insightful comments and valuable suggestions on the dissertation. The discussions with them enriched my vision and deepened my understanding on the research problems.

My sincere thanks also goes to my fellow labmates who accompanied me through out the years, I have learned a lot from each of them. I will always remember the moments we share happiness together as well as the time we go through hardships by helping each other.

I am thankful for my Master's advisor Prof. Tao Jiang, who led me into the research of wireless communication and taught me many research skills. I am also grateful for my high school math teacher Shude Peng and physics teacher Zhengjun Wu, and middle school English teacher Shujun Wei. They not only taught me the knowledge, but also showed me how to become a better individual through hard work.

Last but not the least, I would like to thank my parents for raising me up and providing me good education. Without their tremendous effort and support, I could not come to the United States and complete my Ph.D. study.

At last, this dissertation work is supported in part by the US National Science Foundation under Grant CNS-1320664, Grant CNS-1702957, and Grant CNS-0953513, and through the NSF Broadband Wireless Access and Applications Center (BWAC) site and the Wireless Engineering Research and Education Center (WEREC) at Auburn University. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the NSF.

Table of Contents

List of Figures

Chapter 1

Introduction

With the growing popularity of smart devices and massive Internet of Things (IoT) applications, the fifth generation (5G) wireless communication network is characterized by ubiquitous connection, extremely high traffic demand, and highly complicated network architecture. To support future data-intensive and delay-sensitive applications such as high-quality video streaming, unmanned vehicles, and online gaming, the 5G wireless network is expected to provide reliable wireless services with low delay and high data rate. According to a report by Qualcomm, the 5G wireless network is expected be provide 1000x data rate compared to current cellular systems [1].

To meet with such demand, several promising technologies were proposed as candidates for 5G network. The key technologies for 5G include massive MIMO, small cell network, and millimeter wave (mmWave) communication [2]. A massive MIMO system is characterized by large number of antennas equipped at BS. Due to the law of large numbers, the channel vectors of different users tend to be uncorrelated. Thus, highly directional transmission and efficient spatial multiplexing can be achieved, resulting in improved spectral and energy efficiencies compared to current cellular systems. A small cell network (SCN) consists of multiple low-power and spatially separated small-scale BSs with small coverage areas. With such an architecture, the distance between the transmitter and receiver is greatly reduced, resulting in high signal-to-interference-plus-noise ratio (SINR). The low power of each BS enables more efficient spatial reuse of spectrum, which in turn improves the capacity of the entire network. A mmWave system operates at the higher end of spectrum compared to current cellular systems, which ranges from 30GHz to 300GHz. As a result, a large bandwidth is available (e.g., a 7 GHz

license-free spectrum between 57 GHz and 64 GHz was approved by FCC), which significantly enhances the data rate performance.

Despite these benefits, each technology bears some inherent limitations. For massive MIMO, a key challenge is the high overhead for channel estimation caused by the large number of antennas. Due to this challenge, time division duplex (TDD) systems, which employ channel reciprocity, are considered in most works. However, as the number of users increases, a large proportion of time would be spent on channel estimation, resulting in degraded data rate performance. For small cells network, due to the dense deployment and spectrum reuse, the large number of co-channel transmissions cause interference between different cells. As for mmWave communication, due to the short wavelength, the transmission can be easily blocked by obstacles. Thus, finding alternative links is necessary for reestablishing the link. However, the process of searching for new connections can be difficult and time-consuming due to the narrow beam of mmWave communication. Moreover, how to allocate resource between the reestablished links and existing link remains a challenge. Despite the difference, a common challenge for all technologies is that the system performance degrades as the number of users increase. In this work, we propose efficient designs with particular focus on dealing with large number of users.

Besides the challenges of each individual technology, how to integrate them with proper design remains a challenge. For example, in a massive MIMO heterogeneous network (HetNet), a macrocell BS (MBS) coexists with multiple small cell BSs (SBS). Despite several benefits, a critical issue to be addressed is the interference management. Due to the large dimension of channel matrix, traditional approaches for interference mitigation cannot be directly applied to a massive MIMO HetNet. As another example, in-band full duplex transmission is a promising approach to enhance the spectral efficiency. However, when it is applied in small cells with small coverage area, the interference between a pair of users using the same channel needs to be addressed. Note that, a combination of different technologies can be a good solution to deal with the challenges brought by large number of users. However, a careful configuration is required to fully harvest the benefits. In this work, we also investigate the integration and

coexistence of several combinations of different technologies. We identify the challenges of such integrations and propose effective solutions to enhance the system performance.

The contributions of our work is summarized as follows:

- 1. We consider dynamic BS ON-OFF switching in a massive MIMO HetNet to achieve a good balance between data rate and energy consumption. With the objective of maximizing the energy efficiency, such a problem is formulated as an integer programming. We proposed both centralized and distributed schemes to determine the set of SBSs to be turned off. For the centralized scheme, we first transform the original problem into a convex optimization problem, then a subgradient approach is used to derive the near optimal solution. The distributed schemes are based on a user bidding framework and a wireless operator pricing approach, respectively. Both schemes are proven to achieve a Nash Equilibrium. Simulation results shown that the system energy efficiency can be significantly improved compared to benchmark schemes.

- 2. We consider interference management from the perspective of antenna array processing technique. With non-uniform placement of N antennas (nested array) and a second order processing, a number of $O(N^2)$ degrees of freedom (DoF) can be achieved. Such DoF refers to the number of directions of incoming signals that can be resolved. Based on the direction of arrival information, the desired signals remain while the interference can be nullified using a direction-based beamforming. The critical design issue is how to use the DoF to serve users as well as nullify interference such that the network performance can be optimized. We formulated such problem as an integer programming with objective of maximizing the sum data rate of all users. We proposed centralized and distributed solutions and demonstrated the effectiveness of proposed schemes with simulations.

- 3. We consider a joint frame design, resource allocation, and user association scheme to maximize the sum rate of a massive MIMO HetNet with wireless backhaul. We formulate the problem as an integer programming and propose an iterative solution algorithm. With

adaptive pilot length, i.e., the number of symbols dedicated to pilots in each frame, the proposed scheme outperforms other schemes with fixed frame structures.

- 4. We apply full duplex transmission in a femtocell network and propose efficient solutions to mitigate the interference caused by additional links. We employ a stable roommate matching algorithm to determine the pairing strategy of users and make a selection between half duplex (HD) and full duplex (FD) based on the pairing result. After the duplex selection outcome is determined, we then considered channel and power allocation with the objective improving the sum rate. With the proposed user pairing, adaptive HD/FD selection, and resource allocation, the interference caused by FD transmission is effectively controlled and the system performance is improved compared to the schemes without user pairing and adaptive mode selection.

- 5. We propose a cooperative small cell network architecture to mitigate inter-cell interference and improve the network capacity. The key components include adaptive BS deployment and configuration, dynamic resource allocation, and interference coordination. With efficient spectrum reuse and traffic-aware scheduling, the proposed architecture effectively improves the data rate performance of small cell network when serving large number of users.

- 6. We consider a combination of multiple approaches to enhance the performance of a mmWave system. We designed an adaptive scheme to select the set of users served by D2D relaying and multi-beam reflection. Simulation results demonstrated that significant performance gain can be achieved by adopting a combination of multiple approaches compared to the case with a single approach.

Chapter 2

Dynamic Base Station Sleep Control for Energy Efficient Massive MIMO HetNet

2.1 Introduction

To meet the 1000x mobile data challenge in the near future [1], aggressive spectrum reuse and high spectral efficiency must be achieved to significantly boost the capacity of wireless networks. To this end, *massive MIMO* (Multiple Input Multiple Output) and *small cell* are regarded as two key technologies for emerging 5G wireless systems [3, 4, 7]. Massive MIMO refers to a wireless system with more than 100 antennas equipped at the base station (BS), which serves multiple users with the same time-frequency resource [8]. Due to highly efficient spatial multiplexing, massive MIMO can achieve dramatically improved energy and spectral efficiency over traditional wireless systems [9, 10]. Small cell (or, the notion of *network densification*) is another promising approach for capacity enhancement. With short transmission range and small coverage area, high signal to noise ratio (SNR) and dense spectrum reuse can be achieved, resulting in increased spectral efficiency.

Due to their high potential, the combination of massive MIMO and small cells is expected in future wireless networks, where multiple small cell BS's (SBS) coexist with a macrocell BS (MBS) equipped with a large number of antennas, forming a heterogeneous network (HetNet) with massive MIMO [4–6]. The two technologies are inherently complementary. On one hand, the MBS with massive MIMO has a large number of degrees of freedom (DoF) in the spatial domain, which can be exploited to avoid cross-tier interference. On the other hand, as traffic load grows, the throughput of a massive MIMO system will be limited by factors such as channel estimation overhead and pilot contamination [8]. By offloading some macrocell

users to small cells, the complexity and overhead of channel estimation at the MBS can be greatly reduced, resulting in better performance of macrocell users. Due to these great benefits, massive MIMO HetNet has drawn considerable attention recently [7, 11–16].

However, another advantage of massive MIMO HetNet has not been well considered in the literature, which is its high potential for energy savings. With the rapid growth of wireless traffic and development of data-intensive services, the power consumption of wireless networks has significantly increased, which not only generates more CO_2 emission, but also raises the operating expenditure of wireless operators. As a result, energy saving, or energy efficiency (EE), becomes a rising concern for the design of wireless networks [17]. A few schemes have been proposed to improve the EE of massive MIMO HetNets, such as optimizing the beamforming weights [11] or optimizing user association [14].

In this chapter, we aim to improve the EE of massive MIMO HetNets from the perspective of dynamic ON-OFF switching of BS's. Due to the high potential for spatial-reuse, SBS's are expected to be densely deployed, resulting in considerable energy consumption. As the traffic demand fluctuates over time and space [19, 79], (e.g., a central business district versus a residential area, and daytime versus nighttime), many SBS's are under-utilized for certain periods of a day, and can be turned off to save energy and improve EE. A unique advantage of massive MIMO HetNet is that the MBS can provide good coverage for users that are initially served by the turned-off SBS's. However, as more SBS's are turned off, more users will be served by the MBS. As these users need to send pilot to the MBS, the number of symbols dedicated to the pilot in the transmission frame will be increased, resulting in decreased data rate [16]. Due to this *trade-off*, the SBS ON-OFF switching strategy should be carefully determined to balance the tension between energy saving and data rate performance.

We propose a scheme called BOOST (i.e., BS ON-OFF Switching sTrategy) to maximize the EE of a massive MIMO HetNet, by jointly optimizing BS ON-OFF switching, user association, and power control. We fully consider the special properties of massive MIMO HetNet in problem formulation, develop effective cross-layer optimization algorithms, and provide insights on the solution algorithms.

The joint SBS ON-OFF switching, user association, and power control is formulated as a mixed integer programming problem by taking account of the key design factors. We first propose a centralized solution algorithm, in which an iterative framework with proven convergence is developed With given BS transmit power, the original problem becomes an integer programming problem. To solve the problem with two sets of variables, we relax the integer constraints and transform it into a convex optimization problem. Then, we decompose the relaxed problem into two levels of problems. The lower level problem determines the user association strategy that maximizes the sum rate under given SBS ON-OFF states, the higher level problem updates the SBS ON-OFF strategy based on user association. We derive the optimal solution to the lower level problem with a series of transforms and Lagrangian dual methods. At the higher level problem, we update the SBS ON-OFF states with a subgradient approach. The iteration between the two levels is proven to converge with a guaranteed speed. We then round up the solutions of SBS ON-OFF states to obtain a near-optimal solution to the original problem. With given BS ON-OFF states and user association, the BS transmit power can be optimized with an iterative water-filling approach. To reduce complexity and enhance implementation feasibility, we also propose two distributed schemes based on a user bidding approach and a wireless service provider (WSP) pricing approach, respectively. We show that both games converge to the Nash Equilibrium (NE). The proposed schemes are compared with three benchmarks through simulations, where their performance is validated.

In the remainder of this chapter, we present the system model and problem formulation in Section 2.2. The centralized and distributed schemes are presented in Sections 2.3 and 2.4, respectively. The simulation results are discussed in Section 2.5, the related works are presented in Section 2.6. We conclude this chapter in Section 2.7.

2.2 Problem Formulation

The system considered in this chapter is based on a noncooperative multi-cell network, and we focus on a tagged macrocell. The macrocell is a two-tier HetNet consisting of an MBS with a massive MIMO (indexed by $j = 0$) and J SBS's (indexed by $j = 1, 2, \ldots, J$), which collectively serve K mobile users (indexed by $k = 1, 2, \ldots, K$).

We define binary variables for user association as

$$x_{k,j} \doteq \begin{cases} 1, & \text{user } k \text{ is connected to BS } j \\ 0, & \text{otherwise,} \end{cases} \quad k = 1, 2, \ldots, K,\ j = 0, 1, \ldots, J. \tag{2.1}$$

The MBS is always turned on to guarantee coverage for users in the macrocell. On the other hand, the SBS's can be dynamically switched on or off for energy savings.

The SBS ON-OFF indicator, denoted as y_j, is defined as

$$y_j \doteq \begin{cases} 1, & \text{SBS } j \text{ is turned on} \\ 0, & \text{SBS } j \text{ is turned off,} \end{cases} \quad j = 1, 2, \ldots, J. \tag{2.2}$$

The MBS is equipped with M_0 antennas and adopts linear zero-forcing beamforming. The SBS is equipped with single-antenna and serves multiple users with different time-frequency resources. We consider orthogonal spectrum allocation between the two tiers, where macrocell and small cells operate on different spectrum bands [21, 22, 76].

In the transmission frames of a macrocell user equipment (MUE), a certain number of symbols are dedicated for pilot transmission [8, 23]. Suppose there are N symbols in a frame and B symbols are used as pilot, then the proportion of time for data transmission is $1 - \frac{B}{N}$. According to [23, 24], the total number of users that can served by a massive MIMO system is determined by the number of available uplink (UL) pilots, and B is proportional to the number of MUEs.[1] Specifically, $B = \beta \sum_{k=1}^{K} x_{k,0}$, where β is the *pilot reuse factor* across different macrocells. Without loss of generality, we assume $\beta = 1$. Let $\gamma_{k,0}$ be the average SNR of user k connecting to the MBS. In this chapter, we focus on a widely used model based on zero-forcing precoding. More sophisticated SINR models can be found in [23, 26–28]. The downlink normalized average achievable data rate of user K, when it connects to the MBS, is

[1]Consider a cellular network with frequency use factor 1 as an example. To guarantee the orthogonality between pilots of different UEs, one can either assign mutually orthogonal sequences that span over all available time-frequency blocks to the pilots of UEs, or assign one unique time-frequency block (which should be no larger than a coherence block) to each UE. In both cases, the number of UEs that can be simultaneously served is no larger than $B \cdot N_{\text{smooth}}$, where N_{smooth} is the number of subcarriers in a coherence frequency. In prior works [23–25], the pilot of each user is assigned with one OFDM symbol, then $\sum_{k=1}^{K} x_{k,0} \leq B$. To fully utilize all pilot symbols, we further have $\sum_{k=1}^{K} x_{k,0} = B$.

given as [12, 13, 87]

$$C_{k,0} = \left(1 - \sum_{k=1}^{K} x_{k,0} \left(\frac{T'}{T}\right)\right) \left(\frac{T_u}{T'}\right) \log\left(1 + \frac{M_0 - S_0 + 1}{S_0} \gamma_{k,0}\right), \tag{2.3}$$

where T is the duration of a frame and T' is the interval of a symbol, which corresponds to the time spent to transmit pilot for one user. The interval of a symbol consists of T_u for useful symbol and $T_g = T' - T_u$ for guard interval. M_0 is the number of antennas at MBS, S_0 is the beamforming size, which serves as an upper bound for the number of users that can be simultaneously served by the MBS. Then, $\frac{M_0 - S_0 + 1}{S_0}$ is the antenna array gain of massive MIMO. We assume that the channel state information (CSI) is collected by the MBS via uplink pilot (i.e., a time division duplex (TDD) system), so that the MBS can obtain $\{\gamma_{k,0}\}$.

We assume that the SBS's adopt frequency division multiple access (FDMA), in which the spectrum of SBS j is divided into S_j channels and each of its user is allocated with at least one channel. Thus, the number of users that can be served by SBS j is upper bounded by S_j. In general, proportional fairness is considered as the objective for intra-cell resource allocation. Then equal spectrum allocation is optimal, where each user uses a proportion $\frac{1}{\sum_{k=1}^{K} x_{k,j}}$ of the entire spectrum [31,87]. Let P_j^T be the transmit power of SBS j, then the SINR of user k served by SBS j is $\gamma_{k,j} = \frac{P_j^T \bar{H}_{k,j}}{N_0 + \sum_{l \neq j} P_l^T \bar{H}_{k,l}}$, where $\bar{H}_{k,j}$ is the average channel power gain between BS j and user k [31,87]. Thus, for a user k connecting to SBS j, the downlink normalized achievable data rate of the user can be written as

$$C_{k,j} = \frac{\log(1 + \gamma_{k,j})}{\sum_{k=1}^{K} x_{k,j}} = \frac{R_{k,j}}{\sum_{k=1}^{K} x_{k,j}}, \ j = 1, 2, \ldots, J, \tag{2.4}$$

where $R_{k,j} \doteq \log(1 + \gamma_{k,j})$, $j = 1, 2, \ldots, J$.

In this chapter, we consider three time scales: the period of BS on-off switching, T_1; the period of user association and power control update, T_2; and the period of CSI acquisition, T_3. Since it is infeasible to turn on/off a BS frequently, T_1 is much large than T_2. Before the update user association, the time averaged SNR or SINR of each user is measured within an interval of T_3 to offset the effect of fast fading.

The power consumption model of HetNets is studied in [33]. The power consumption of a BS consists of a static part and a dynamic part. The static part is the power required for the operation of a BS once it is turned on, e.g., used by the cooling system, power amplifier, and baseband units. The dynamic part is mainly used by the radio frequency unit. Thus, the power consumption of each BS is given as $P_j = P_j^S + P_j^T$, $j = 0, 1, \ldots, J$, where P_j^S is the constant power consumption when a BS is turned on, P_j^T is the transmit power. Then, the total power consumption of the HetNet is $P_0 + \sum_{j=1}^{J} y_j P_j$.

In this chapter, we aim to dynamically switch off under-utilized SBS's and maximize the EE of a HetNet with massive MIMO. The EE of a HetNet, defined by the sum rate divided by the total power, has been widely considered as the objective function in prior works [34, 35]. In particular, such objective was used in a recent study on the achievable EE of massive MIMO HetNet [24]. Theoretically, the EE can be maximized if we turn on an SBS whenever there is a user to be served and allocate all channels to the user, and then turn off the SBS after the transmission is finished. However, this results in frequent on-off switching of BS, which is not practical since the on-off switching is time-consuming and introduces additional power consumption. As the SBS on-off switching is performed at a much larger timescale than that of user association, an SBS is expected to serve a certain number of users during its active period. Due to this fact, a BS is turned on when the traffic load or user requests exceed a threshold in many previous works such as in [37, 65]. On the other hand, if we directly use EE as the objective, the aggregated data rate of users in a small cell would remain at a high level even when there are only a small number of users in the small cell, since each user is allocated with a large bandwidth. Then, an SBS would not be turned off even if its traffic is low. To this end, we adjust the expression of EE by replacing $C_{k,j}$ with its worst case value, $\tilde{C}_{k,j} = \frac{\log\left(1+\gamma_{k,j}\right)}{S_j}$.

Let $\mathbf{x}$, $\mathbf{y}$ and $\mathbf{P}_T$ denote the $\{x_{k,j}\}$ matrix, the $\{y_j\}$ vector, and the $\{P_j^T\}$ vector, respectively. The problem can be formulated as

$$\mathbf{P1}: \max_{\{\mathbf{x},\mathbf{y},\mathbf{P_T}\}} \frac{\sum_{k=1}^{K} x_{k,0} C_{k,0} + \sum_{k=1}^{K}\sum_{j=1}^{J} x_{k,j}\tilde{C}_{k,j}}{P_0 + \sum_{j=1}^{J} y_j P_j} \tag{2.5}$$

$$\text{s.t.: } \sum_{j=0}^{J} x_{k,j} \leq 1,\ k = 1, 2, \ldots, K \tag{2.6}$$

$$\sum_{k=1}^{K} x_{k,j} \leq S_j,\ j = 0, 1, \ldots, J \tag{2.7}$$

$$x_{k,j} \leq y_j,\ k = 1, 2, \ldots, K,\ j = 1, 2, \ldots, J \tag{2.8}$$

$$P_j^T \leq P_{\max}^T,\ j = 1, \ldots, J \tag{2.9}$$

$$x_{k,j} \in \{0, 1\},\ k = 1, 2, \ldots, K,\ j = 0, 1, \ldots, J \tag{2.10}$$

$$y_j \in \{0, 1\},\ j = 1, 2, \ldots, J. \tag{2.11}$$

In problem **P1**, constraint (2.6) is due to the fact that each user can connect to at most one BS; constraint (2.7) enforces the upper bound on the number of users that can be served by BS j; and constraint (2.8) is because users can connect to SBS j only when it is turned on. $P_{\max}^T$ is the maximum transmit power of an SBS.

2.3 Centralized Solution

Usually small cells are deployed by the operator and can use the X2 interface, which is the interface used between eNodeBs [40], to communicate with each other as well as the MBS. A centralized algorithm can be useful in this context to coordinate their operations. In this section, we solve the formulated problem with a centralized scheme and show that near-optimal solution can be achieved. Since Problem **P1** is a mixed integer non-convex problem with 3 sets of coupled variables, we propose an iteratively approach to solve $\{\mathbf{x}, \mathbf{y}\}$ and $\mathbf{P}_T$. With given $\mathbf{P}_T$, we obtain the near optimal $\mathbf{y}$ with a subgradient approach and derive the optimal $\mathbf{x}$ with given $\mathbf{y}$. With given $\mathbf{x}$ and $\mathbf{y}$, we derive the power control solution $\mathbf{P}_T$ that mitigates mutual interference. We show that the iteration between $\{\mathbf{x}, \mathbf{y}\}$ and $\mathbf{P}_T$ converges.

2.3.1 Near Optimal BS ON-OFF Switching with Given Transmit Power: A Subgradient Approach

With given $\mathbf{P}_T$, Problem **P1** becomes an integer programming problem, which is still NP-hard. To develop an effective solution algorithm, we relax the integer constraints by allowing $x_{k,j}$ and y_j to take values in $[0, 1]$. However, the objective function of the relaxed problem of **P1** is non-convex, the global optimum is not achievable. To this end, we define substitution variables $\tilde{y}_j = \log y_j$ and transform the objective function into an equivalent form. Then, we have the following problem.

$$\mathbf{P2}: \max_{\{\mathbf{x},\tilde{\mathbf{y}}\}} \left\{ \log\left(\sum_{k=1}^{K} x_{k,0} C_{k,0} + \sum_{k=1}^{K} \sum_{j=1}^{J} x_{k,j} \tilde{C}_{k,j} \right) - \log\left(P_0 + \sum_{j=1}^{J} e^{\tilde{y}_j} P_j \right) \right\} \tag{2.12}$$

$$\text{s.t.:} \quad \sum_{j=0}^{J} x_{k,j} \leq 1, \ k = 1, 2, \ldots, K \tag{2.13}$$

$$\sum_{k=1}^{K} x_{k,j} \leq S_j, \ j = 0, 1, \ldots, J \tag{2.14}$$

$$\log x_{k,j} \leq \tilde{y}_j, \ k = 1, 2, \ldots, K, \ j = 1, 2, \ldots, J \tag{2.15}$$

$$0 \leq x_{k,j} \leq 1, \ k = 1, 2, \ldots, K, \ j = 0, 1, \ldots, J \tag{2.16}$$

$$\tilde{y}_j \leq 0, \ j = 1, 2, \ldots, J. \tag{2.17}$$

We first show that problem **P2** is a convex problem so that dual methods can be applied.

Lemma 1 *Problem* **P2** *is a convex optimization problem.*

Proof: The objective function of problem **P2** has two parts. In the first part, we first consider the sum rate expression inside the log function, $\sum_{k=1}^{K} x_{k,0} C_{k,0} + \sum_{k=1}^{K} \sum_{j=1}^{J} x_{k,j} \tilde{C}_{k,j}$. It is a combination of a linear function of $\mathbf{x}$ and the term $E \doteq -\frac{T_u}{T} \left(\sum_{k=1}^{K} x_{k,0} \right) \left(\sum_{k=1}^{K} x_{k,0} R_{k,0} \right)$,

where $R_{k,0} = \log(1 + \frac{M_0 - S_0 + 1}{S_0}\gamma_{k,0})$. The Hessian of E is given by

$$\mathbf{H}_{K\times K} = -\frac{T_u}{T}\begin{pmatrix} 2R_{1,0} & R_{1,0}+R_{2,0} & \cdots & R_{1,0}+R_{K,0} \\ R_{1,0}+R_{2,0} & 2R_{2,0} & \cdots & R_{2,0}+R_{K,0} \\ \vdots & \vdots & \ddots & \vdots \\ R_{1,0}+R_{K,0} & R_{2,0}+R_{K,0} & \cdots & 2R_{K,0} \end{pmatrix}.$$

Let $\mathbf{z} = [z_1, z_2, \ldots, z_k]^T$ be an arbitrary non-zero vector. We have $\mathbf{z}^T\mathbf{H}\mathbf{z} \overset{(a)}{<} -\frac{2T'}{T}\left[\sum_{k=1}^{K} z_k^2 R_{k,0} + \sum_{k=1}^{K}\sum_{k'\neq k} z_k z_{k'}\left(2\sqrt{R_{k,0}R_{k',0}}\right)\right] = -\frac{2T'}{T}\left(\sum_{k=1}^{K} z_k\sqrt{R_{k,0}}\right)^2 < 0$, where inequality (a) results from the fact that for two positive numbers, $m + n \geq 2\sqrt{mn}$ and the equality holds when $m = n$.

We conclude that E is a concave function. Then, $\sum_{k=1}^{K} x_{k,0}C_{k,0} + \sum_{k=1}^{K}\sum_{j=1}^{J} x_{k,j}\tilde{C}_{k,j}$ is also concave. As $\log(\cdot)$ is a concave function, the first part of the objective function of problem **P2** is a concave function.

The second part, given as $-\log\left(P_0 + \sum_{j=1}^{J} e^{\tilde{y}_j} P_j\right)$, is a log-sum-exp, which is concave according to [41]. Therefore, the objective function is concave. Constraint $\log x_{k,j} - \tilde{y}_j \leq 0$ is a concave function, the other constraints are linear functions. Thus, problem **P2** is a convex optimization problem. ■

In problem **P2**, the decision variables $x_{k,j}$ and $\tilde{y}_j$ are coupled in the constraints, which are difficult to handle directly. Besides, the objective function includes a weighted sum of quadratic expressions, which is highly complex. To obtain the optimal solution of problem **P2**, we introduce an auxiliary variable $Q_0 \doteq \sum_{k=1}^{K} x_{k,0}$. Then, both Q_0 and $\tilde{y}_j$ are coupling variables with $x_{k,j}$. To decouple the variables, we decompose problem **P2** into two levels of subproblems. At the lower-level subproblem, we find the optimal solution of $\mathbf{x}$ for given values of $\tilde{\mathbf{y}}$ and Q_0. Based on the solution at the lower-level subproblem, we obtain the optimal values of $\tilde{\mathbf{y}}$ and Q_0 at the higher-level subproblem through a subgradient approach.

Lower-level of Problem **P2**: The Optimal Solution of $\mathbf{x}$ with Given $\tilde{\mathbf{y}}$ and Q_0

For given values of $\tilde{\mathbf{y}}$ and Q_0, the lower-level subproblem of problem **P2** is given as

$$\mathbf{P3}: \max_{\{\mathbf{x}\}} \left\{ \log\left(\sum_{k=1}^{K} x_{k,0} C_{k,0} + \sum_{k=1}^{K} \sum_{j=1}^{J} x_{k,j} \tilde{C}_{k,j} \right) - \log\left(P_0 + \sum_{j=1}^{J} e^{\tilde{y}_j} P_j \right) \right\} \tag{2.18}$$

$$\text{s.t.: } (2.13) - (2.17) \text{ and}$$

$$\sum_{k=1}^{K} x_{k,0} = Q_0. \tag{2.19}$$

We take a partial relaxation on the constraints on Q_0 and $\tilde{y}_j$, i.e., (2.17) and (2.19). The dual problem of **P3** is given by

$$\textbf{P3-Dual:} \quad \min_{\{\boldsymbol{\lambda}, \mu\}} \; g(\boldsymbol{\lambda}, \mu), \tag{2.20}$$

where $\boldsymbol{\lambda}$ and μ are the Lagrangian multipliers for constraints (2.15) and (2.19), respectively; and $g(\boldsymbol{\lambda}, \mu)$ is given by

$$g(\boldsymbol{\lambda}, \mu) = \max_{\{\mathbf{x}\}} \left\{ \log\left(\sum_{k=1}^{K} x_{k,0} C_{k,0} + \sum_{k=1}^{K} \sum_{j=1}^{J} x_{k,j} \tilde{C}_{k,j} \right) - \log\left(P_0 + \sum_{j=1}^{J} e^{\tilde{y}_j} P_j \right) + \sum_{k=1}^{K} \sum_{j=1}^{J} \lambda_{k,j} \left(\tilde{y}_j - \log x_{k,j} \right) + \mu \left(Q_0 - \sum_{k=1}^{K} x_{k,0} \right) \right\}.$$

The optimal solution of **P3-Dual** can be obtained with the following subgradient method.

$$\begin{cases} \lambda_{k,j}^{[t+1]} = \left[\lambda_{k,j}^{[t]} + \frac{g(\boldsymbol{\lambda}^{[t]}, \mu^{[t]}) - g(\boldsymbol{\lambda}^{*}, \mu^{[t]})}{\left\| \boldsymbol{\delta}_{\lambda}^{[t]} \right\|^2} \left(\log x_{k,j}^{[t]} - \tilde{y}_j^{[t]} \right) \right]^{+}, \\ \qquad\qquad\qquad\qquad \forall\, k, j \\ \mu^{[t+1]} = \mu^{[t]} + \frac{g\left(\boldsymbol{\lambda}^{[t]}, \mu^{[t]}\right) - g\left(\boldsymbol{\lambda}^{[t]}, \mu^{*}\right)}{|Q_0{}^{[t]} - \sum_{k=1}^{K} x_{k,0}^{[t]}|} \left(\sum_{k=1}^{K} x_{k,0}^{[t]} - Q_0^{[t]} \right), \end{cases} \tag{2.21}$$

where $[z]^{+} \doteq \max\{0, z\}$, and t is the index of iteration. $\boldsymbol{\delta}_{\lambda}^{[t]}$ is the vector of gradients of $\{\lambda_{k,j}\}$ given as $\left[\tilde{y}_1^{[t]} - \log x_{1,1}^{[t]}, ..., \tilde{y}_J^{[t]} - \log x_{K,J}^{[t]} \right]^{T}$. Since $\boldsymbol{\lambda}^{*}$ and μ^{*} are unknown before solving the

problem, we use the mean of objective values of the primal and dual problems as an estimate for $g(\boldsymbol{\lambda}^*, \mu^{[t]})$ and $g(\boldsymbol{\lambda}^{[t]}, \mu^*)$ [31]. $g(\boldsymbol{\lambda}, \mu)$ can be obtained by solving the following problem

$$\mathbf{P4}: \max_{\{\mathbf{x}\}} \mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}, \mu) \quad \text{s.t. } (2.13), (2.14), \text{and } (2.16), \tag{2.22}$$

where $\mathcal{L}(\cdot)$ is the *Lagrangian function*. With given $\lambda_{k,j}$, μ, and Q_0, problem **P4** is a standard convex optimization problem which can be solved using KKT conditions.

Lemma 2 *The sequence $g(\boldsymbol{\lambda}^{[t]}, \mu^{[t]})$ converges to $g(\boldsymbol{\lambda}^*, \mu^*)$ with a speed faster than $\left\{1/\sqrt{t}\right\}$ as t goes to infinity.*

Proof: The vector form of (2.21) is given as $\boldsymbol{\lambda}^{[t+1]} = \left[\boldsymbol{\lambda}^{[t]} + \frac{g(\boldsymbol{\lambda}^{[t]}, \mu^{[t]}) - g(\boldsymbol{\lambda}^*, \mu^{[t]})}{\left\|\boldsymbol{\delta}_\lambda^{[t]}\right\|^2} \boldsymbol{\delta}_\lambda^{[t]}\right]^+$. Consider the optimality gap of $\boldsymbol{\lambda}$, we have

$$\begin{aligned}
&\|\boldsymbol{\lambda}^{[t+1]} - \boldsymbol{\lambda}^*\|^2 \\
\leq &\left\| \boldsymbol{\lambda}^{[t]} + \frac{g(\boldsymbol{\lambda}^{[t]}, \mu^{[t]}) - g(\boldsymbol{\lambda}^*, \mu^{[t]})}{\|\boldsymbol{\delta}_\lambda^{[t]}\|^2} \boldsymbol{\delta}_\lambda^{[t]} - \boldsymbol{\lambda}^* \right\|^2 + \left(\frac{g(\boldsymbol{\lambda}^{[t]}, \mu^{[t]}) - g(\boldsymbol{\lambda}^*, \mu^{[t]})}{\|\boldsymbol{\delta}_\lambda^{[t]}\|^2} \right)^2 \|\boldsymbol{\delta}_\lambda^{[t]}\|^2 \\
\overset{(a)}{\leq} &\|\boldsymbol{\lambda}^{[t]} - \boldsymbol{\lambda}^*\|^2 - 2\frac{\left(g(\boldsymbol{\lambda}^{[t]}, \mu^{[t]}) - g(\boldsymbol{\lambda}^*, \mu^{[t]})\right)^2}{\|\boldsymbol{\delta}_\lambda^{[t]}\|^2} + \left(\frac{g(\boldsymbol{\lambda}^{[t]}, \mu^{[t]}) - g(\boldsymbol{\lambda}^*, \mu^{[t]})}{\|\boldsymbol{\delta}_\lambda^{[t]}\|^2} \right)^2 \|\boldsymbol{\delta}_\lambda^{[t]}\|^2 \\
\leq &\|\boldsymbol{\lambda}^{[t]} - \boldsymbol{\lambda}^*\|^2 - \frac{\left(g(\boldsymbol{\lambda}^{[t]}, \mu^{[t]}) - g(\boldsymbol{\lambda}^*, \mu^{[t]})\right)^2}{\hat{\boldsymbol{\delta}}_\lambda^2},
\end{aligned}$$

where inequality (a) is due to convexity of problem **P3-dual**, $\hat{\boldsymbol{\delta}}_\lambda$ is an upper bound for $\boldsymbol{\delta}_\lambda^{[t]}$. Since $\lim_{t\to\infty} \boldsymbol{\lambda}^{[t+1]} = \lim_{t\to\infty} \boldsymbol{\lambda}^{[t]}$, it follows that $\lim_{t\to\infty} g(\boldsymbol{\lambda}^{[t]}, \mu^{[t]}) = g(\boldsymbol{\lambda}^*, \mu^{[t]})$. Summing the above inequality over t, we have

$$\sum_{t=1}^{\infty} \left(g(\boldsymbol{\lambda}^{[t]}, \mu^{[t]}) - g(\boldsymbol{\lambda}^*, \mu^{[t]})\right)^2 \leq \hat{\boldsymbol{\delta}}_\lambda^2 \left\|\boldsymbol{\lambda}^{[1]} - \boldsymbol{\lambda}^*\right\|^2. \tag{2.23}$$

Suppose $\lim_{t\to\infty} \left(g(\boldsymbol{\lambda}^{[t]}, \mu^{[t]}) - g(\boldsymbol{\lambda}^*, \mu^{[t]})\right) \sqrt{t} > 0$ for contradiction. There must be a sufficiently large t' and a positive number ξ such that $\left(g(\boldsymbol{\lambda}^{[t]}, \mu^{[t]}) - g(\boldsymbol{\lambda}^*, \mu^{[t]})\right) \sqrt{t} \geq \xi, \forall t \geq t'$.

Taking the square sum from t' to ∞, we have

$$\sum_{t=t'}^{\infty} \Big(g(\boldsymbol{\lambda}^{[t]}, \mu^{[t]}) - g(\boldsymbol{\lambda}^{*}, \mu^{[t]}) \Big)^2 \geq \xi^2 \sum_{t=t'}^{\infty} \frac{1}{t} = \infty. \tag{2.24}$$

It is obvious that (2.24) contradicts with (2.23). Thus, the assumption does not hold and we have

$$\lim_{t \to \infty} \frac{g(\boldsymbol{\lambda}^{[t]}, \mu^{[t]}) - g(\boldsymbol{\lambda}^{*}, \mu^{[t]})}{1/\sqrt{t}} = 0, \tag{2.25}$$

this indicates that the convergence speed of the sequence $g(\boldsymbol{\lambda}^{[t]}, \mu^{[t]})$ is faster than that of $1/\sqrt{t}$.

Note that, the updates of $\boldsymbol{\lambda}$ and μ are independent, and are performed in parallel. Applying the same analysis to μ, we conclude that $g(\boldsymbol{\lambda}^{[t]}, \mu^{[t]})$ converges to $g(\boldsymbol{\lambda}^{*}, \mu^{*})$ with a speed faster than that of $1/\sqrt{t}$ as well. ■

Higher-level of Problem **P2**: The Optimal Solution of $\tilde{\mathbf{y}}$ and Q_0

We first show that the *duality gap* between the lower level subproblem **P3** and its dual, problem **P3-Dual**, is zero.

Lemma 3 *Strong duality holds for problem* **P3**.

Proof: It can be easily verified that there exists a feasible $\mathbf{x}$ such that all linear constraints are satisfied while inequalities hold(2.15), the problem is strictly feasible. Thus, the Slater's condition is satisfied and strong duality holds. ■

Let $f(\mathbf{x})$ be the objective function of problem **P3** for a given $\mathbf{x}$. In the higher-level subproblem of problem **P2**, we find the optimal $\tilde{\mathbf{y}}$ and Q_0 by solving the following problem.

$$\mathbf{P5}: \max_{\{\tilde{\mathbf{y}}, Q_0\}} f(\mathbf{x}(\tilde{\mathbf{y}}, Q_0)). \tag{2.26}$$

Lemma 4 *Problem* **P5** *can be solved with the following subgradient method.*

$$\begin{cases} Q_0^{[t+1]} = Q_0^{[t]} + \frac{f(\tilde{\mathbf{y}}^{[t]}, Q_0^{[t]}) - f(\tilde{\mathbf{y}}^{[t]}, Q_0^*)}{\left|\gamma^{[t]}\right|}\gamma^{[t]} \\ \tilde{\mathbf{y}}^{[t+1]} = \tilde{\mathbf{y}}^{[t]} + \frac{f(\tilde{\mathbf{y}}^{[t]}, Q_0^{[t]}) - f(\tilde{\mathbf{y}}^*, Q_0^{[t]})}{\left\|\boldsymbol{\nu}^{[t]}\right\|^2}\boldsymbol{\nu}^{[t]}, \end{cases} \tag{2.27}$$

where $\boldsymbol{\nu}^{[t]} = \left[\sum_{k=1}^{K} \lambda_{k,1}^{*}{}^{[t]} - \frac{P_1 e^{\tilde{y}_1{}^{[t]}}}{P_0 + \sum_{j=1}^{J} P_j e^{\tilde{y}_j{}^{[t]}}}, ..., \sum_{k=1}^{K} \lambda_{k,J}^{*}{}^{[t]} - \frac{P_J e^{\tilde{y}_J{}^{[t]}}}{P_0 + \sum_{j=1}^{J} P_j e^{\tilde{y}_j{}^{[t]}}}\right]^T$, *and* $\gamma^{[t]} = \mu^{*[t]} - \frac{\frac{T'}{T}\sum_{k=1}^{K} x_{k,0}^{[t]} R_{k,0}}{\sum_{k=1}^{K}\sum_{j=0}^{J} x_{k,j}^{[t]} C_{k,j}}$.

Note that, Q_0 *has to be an integer no greater than* S_0 *due to constraint (2.7). After* Q_0 *converges, the final value of* Q_0 *is rounded up to the integer which achieves a greater value of objective function and no larger than* S_0*. The update given by (2.27) should project to the feasible regions of* Q_0 *and* $\tilde{\mathbf{y}}$*, and terminates if the boundary values are obtained.*

Proof: In (2.27), $\tilde{\mathbf{y}}$ and Q_0 are updated in parallel and independently. We first show that Q_0 can updated with the subgradient approach given in (2.27).

Let $\mathbf{x}^*(Q_0{}')$ be the optimal solution to problem **P4** for a given value of $Q_0{}' = \sum_{k=1}^{K} x_{k,0}^*$, and $f^*(Q_0{}')$ be the optimal objective value with solution $\mathbf{x}^*(Q_0{}')$. Consider another feasible solution $\mathbf{x}$ to problem **P4** with $Q_0 = \sum_{k=1}^{K} x_{k,0}$, the following equalities and inequalities hold.

$$\begin{aligned} f^*(Q_0{}') &\overset{(a)}{=} \mathcal{L}\left(\mathbf{x}^*(Q_0{}'), \boldsymbol{\lambda}^*(Q_0{}'), \mu^*(Q_0{}')\right) \overset{(b)}{\geq} \mathcal{L}\left(\mathbf{x}(Q_0{}'), \boldsymbol{\lambda}^*(Q_0{}'), \mu^*(Q_0{}')\right) \\ &\overset{(c)}{\geq} f(\mathbf{x}(Q_0)) - \frac{\frac{T'}{T}\sum_{k=1}^{K} x_{k,0} R_{k,0}}{\sum_{k=1}^{K} x_{k,0} C_{k,0}}(Q_0{}' - Q_0) \\ &\quad + \sum_{k=1}^{K}\sum_{j=1}^{J} \lambda_{k,j}^*(\tilde{y}_j - \log x_{k,j}) + \mu^*\left(Q_0{}' - \sum_{k=1}^{K} x_{k,0}\right) \\ &\overset{(d)}{\geq} f\left(\mathbf{x}(Q_0)\right) + \left(\mu^* - \frac{T'}{T}\sum_{k=1}^{K} x_{k,0} R_{k,0} / \sum_{k=1}^{K} x_{k,0} C_{k,0}\right)(Q_0{}' - Q_0), \end{aligned}$$

where equality (a) is due to strong duality, inequality (b) is due to the optimality of $\mathbf{x}^*$, inequality (c) is because $-\frac{T'}{T}\sum_{k=1}^{K} x_{k,0} R_{k,0} / \sum_{k=1}^{K} x_{k,0} C_{k,0}$ is a gradient of $f(\mathbf{x}(Q_0))$ as a function of

Q_0 with given $\mathbf{x}$, and inequality (d) is due to the constraints of problem **P4** and the nonnegativity of $\boldsymbol{\lambda}$. Note that, (d) holds for any $\mathbf{x}$ such that $\sum_{k=1}^{K} x_{k,0} = Q_0$.

In particular, we have

$$\begin{aligned} f^*({Q_0}') \geq & \max_{\{\mathbf{x}|\sum_{k=1}^{K} x_{k,0}=Q_0\}} \{f(\mathbf{x}) \\ & + \left(\mu^* - \frac{T'}{T}\frac{\sum_{k=1}^{K} x_{k,0}R_{k,0}}{\sum_{k=1}^{K} x_{k,0}C_{k,0}}\right)({Q_0}' - Q_0)\Bigg\} \\ = & f^*(Q_0) + \left(\mu^* - \frac{T'}{T}\frac{\sum_{k=1}^{K} x_{k,0}R_{k,0}}{\sum_{k=1}^{K} x_{k,0}C_{k,0}}\right)({Q_0}' - Q_0). \end{aligned} \tag{2.28}$$

It follows (2.28) that $f^*(Q_0) \leq f^*({Q_0}') + \left(\mu^*({Q_0}') - \frac{\frac{T'}{T}\sum_{k=1}^{K} x_{k,0}({Q_0}')R_{k,0}}{\sum_{k=1}^{K} x_{k,0}({Q_0}')C_{k,0}}\right)(Q_0 - {Q_0}')$. By definition, $\mu^*({Q_0}') - \frac{\frac{T'}{T}\sum_{k=1}^{K} x_{k,0}({Q_0}')R_{k,0}}{\sum_{k=1}^{K} x_{k,0}({Q_0}')C_{k,0}}$ is a subgradient of $f^*(Q_0)$. Therefore Q_0 can be updated with the approach given in (2.27).

Then, we consider the update of $\tilde{\mathbf{y}}$. The objective function of problem **P2** has two parts. The first part, $\log(\sum_{k=1}^{K}\sum_{j=0}^{J} x_{k,j}C_{k,j})$, is an indirect function of $\tilde{\mathbf{y}}$; the second part, given as $-\log(P_0 + \sum_{j=1}^{J} e^{\tilde{y}_j}P_j)$, is a differentiable function of $\tilde{\mathbf{y}}$. Then, a primal decomposition can be applied to maximize the two parts separately.

Denote $D^*(\tilde{\mathbf{y}})$ as the optimal value of the *first* part with given $\tilde{\mathbf{y}}$. Let $\mathbf{x}^*(\tilde{\mathbf{y}}')$ be the optimal solution to problem **P2** for a given $\tilde{\mathbf{y}}'$ and $\mathbf{x}$ be another feasible solution for given $\tilde{\mathbf{y}}$. Then, we have the following inequalities and equalities.

$$\begin{aligned} & D^*(\tilde{\mathbf{y}}') = D(\mathbf{x}^*(\tilde{\mathbf{y}}')) = \mathcal{L}(\mathbf{x}^*, \boldsymbol{\lambda}^*(\tilde{\mathbf{y}}')) \geq \mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}^*(\tilde{\mathbf{y}}')) \\ = & D(\mathbf{x}) + \sum_{k=1}^{K} \boldsymbol{\lambda}_k^*(\tilde{\mathbf{y}}')(\tilde{\mathbf{y}}' - \boldsymbol{\varphi}_k) \\ = & D(\mathbf{x}) + \sum_{k=1}^{K} \boldsymbol{\lambda}_k^*(\tilde{\mathbf{y}}')(\tilde{\mathbf{y}} - \boldsymbol{\varphi}_k) + \sum_{k=1}^{K} \boldsymbol{\lambda}_k^*(\tilde{\mathbf{y}}')(\tilde{\mathbf{y}}' - \tilde{\mathbf{y}}) \\ \geq & D(\mathbf{x}) + \sum_{k=1}^{K} \boldsymbol{\lambda}_k^*(\tilde{\mathbf{y}}')(\tilde{\mathbf{y}}' - \tilde{\mathbf{y}}), \end{aligned} \tag{2.29}$$

where $\boldsymbol{\varphi}_k = [\log x_{k,1}, \log x_{k,2}, \ldots, \log x_{k,J}]^T$ and $\boldsymbol{\lambda}_k^*(\mathbf{y}')$ is the kth row of $\boldsymbol{\lambda}^*(\mathbf{y}')$. In particular, we have

$$D^*(\tilde{\mathbf{y}}') \geq \max_{\{\mathbf{x}|\boldsymbol{\varphi}\leq\tilde{\mathbf{y}}\}} \left\{ D(\mathbf{x}) + \sum_{k=1}^{K} \boldsymbol{\lambda}_k^*(\tilde{\mathbf{y}}' - \mathbf{y}) \right\} = D^*(\tilde{\mathbf{y}}) + \sum_{\mathbf{k=1}}^{\mathbf{K}} \boldsymbol{\lambda}_{\mathbf{k}}^*(\tilde{\mathbf{y}}' - \tilde{\mathbf{y}}). \tag{2.30}$$

Thus, $\left[\sum_{k=1}^{K} {\lambda_{k,1}^*}^{[t]}, ..., \sum_{k=1}^{K} {\lambda_{k,J}^*}^{[t]}\right]^T$ is a subgradient of $\tilde{\mathbf{y}}$ as a function of $D^*(\tilde{\mathbf{y}})$.

The *second* part of the objective function of problem **P2** is a differentiable concave function. We have

$$-\log\left(P_0 + \sum_{j=1}^{J} e^{\tilde{y}_j} P_j\right) \leq -\log\left(P_0 + \sum_{j=1}^{J} e^{\tilde{y}'_j} P_j\right) + \sum_{j=1}^{J} \frac{e^{\tilde{y}_j} P_j (\tilde{y}'_j - \tilde{y}_j)}{P_0 + \sum_{j=1}^{J} e^{\tilde{y}_j} P_j}. \tag{2.31}$$

According to the principles of primal decomposition, $\tilde{\mathbf{y}}$ can be updated by combining (2.30) and (2.31) to achieve its optimal value. Thus, $\boldsymbol{\nu}$ is a subgradient of the objective function of problem **P2** as a function of $\tilde{\mathbf{y}}$. We conclude that problem **P5** can be solved with (2.27). ■

Using the same approach for $\boldsymbol{\lambda}$ and μ, we can also prove that $\tilde{\mathbf{y}}$ and Q_0 converge faster than the sequence $\{1/\sqrt{t}\}$.

Theorem 1 *The complexity of solving problem* **P2** *is upper bounded by* $1/(\varepsilon_1^2\varepsilon_2^2\varepsilon_3^2)$*, where* ε_1 *is the threshold of convergence for* $\tilde{\mathbf{y}}$ *and* Q_0*;* ε_2 *is the threshold of convergence for* $\boldsymbol{\lambda}$ *and* μ*; and* ε_3 *is the threshold of convergence for the dual variables of problem* **P4**.

Proof: According to Lemma 2 and (2.25), for a sufficiently large t and a sufficiently small ε_2, the optimality gap is smaller than $1/\sqrt{t}$. Thus, when $1/\sqrt{t} > \varepsilon_2$, the optimality gap, $g(\boldsymbol{\lambda}^{[t]}, \mu^{[t]}) - g(\boldsymbol{\lambda}^*, \mu^*)$, is guaranteed to be smaller than ε_2. Consequently, we have $t < 1/\varepsilon_2^2$, it takes less than $1/\varepsilon_2^2$ steps for the sequence $g(\boldsymbol{\lambda}^{[t]}, \mu^{[t]})$ to achieve a optimality gap that is less than ε_2. In the same way, the number of updates for $\{\tilde{\mathbf{y}}, \mathbf{Q_0}\}$ and the dual variables in problem **P4** are upper bounded by $1/\varepsilon_1^2$ and $1/\varepsilon_3^2$, respectively.

In the proposed scheme, each update of $\tilde{\mathbf{y}}$ and Q_0 requires a set of optimal $\boldsymbol{\lambda}$ and μ under the current $\tilde{\mathbf{y}}$ and Q_0; each update of $\boldsymbol{\lambda}$ and μ requires the solution of problem **P4** under the

current $\boldsymbol{\lambda}$ and μ. Thus, the total number of variable updates is upper bounded by $1/(\varepsilon_1^2\varepsilon_2^2\varepsilon_3^2)$. Therefore, the complexity of solving problem **P2** is upper bounded by $1/(\varepsilon_1^2\varepsilon_2^2\varepsilon_3^2)$. ■

Near Optimal Solution of SBS ON-OFF States $\mathbf{y}$

With the optimal solution of $\tilde{\mathbf{y}}$ for problem **P2**, the optimal $\mathbf{y}$ with $0 \leq y_j \leq 1$ $j = 1, 2, ..., J$ can be obtained. However, it is highly possible that not all the values of $\{y_j\}$ are 0-1 integers. To determine the actual SBS ON-OFF states, we develop a heuristic scheme to obtain a near optimal integer solutions of $\mathbf{y}$.

Consider the update of $\tilde{y}_j^{[t]}$, the subgradient is given as $\sum_{k=1}^{K} {\lambda_{k,j}^*}^{[t]} - \frac{P_j e^{\tilde{y}_j\,[t]}}{P_0+\sum_{j=1}^{J} P_j e^{\tilde{y}_j^{[t]}}}$. The first part can be interpreted as a measure for the sum rate of all users served by SBS j with the current value of $\tilde{y}_j$. This is because the value of $\lambda_{k,j}$ is determined by the value of $x_{k,j}$ as indicated in (2.21), and a large $x_{k,j}$ indicates that a large rate can be achieved if user k connects to BS j. The second part is a measure of the power consumption of SBS j. Thus, an SBS with large value of $\tilde{y}_j$ has a better capability of providing high sum rate with relatively small power, i.e., being more energy efficient.

Based on this observation, we propose a heuristic scheme to find the set of SBS's to be turned on that achieve the highest EE. Denote the number of SBS's that are turned on as κ, which is an integer between 0 and J. For a given κ, we choose to turn on the first κ SBS's with the largest values of $\tilde{y}_j$, i.e., set $y_j = 1$ for these SBS's and $y_j = 0$ for other SBS's. Then, we evaluate the system EE under different values of κ, and find the one with the largest value. Note that, to calculate the EE, we need to acquire the user association strategy under integral $\mathbf{y}$, which will be discussed in the following part. Once the optimal κ is obtained, the corresponding set of SBS's that are turned on is determined, we have a near-optimal solution of $\mathbf{y}$. The procedure is summarized in Algorithm 1.

The solution produced by Algorithm 1 is expected to be very close to the optimal solution, or be the optimal solution for a network that is not ultra-dense. In such a network, the overlap of coverages of different SBS's is small. Thus, for most users, there is one SBS that can provide a much higher data rate than other SBS's. The mutual impact of ON-OFF states of different SBS's is very limited. As a result, the case of partial user association, $0 < x_{k,j} < 1$, would

Algorithm 1: Centralized BS ON-OFF Switching Strategy

Initialize $Q_0, \tilde{\mathbf{y}}, \boldsymbol{\lambda}$, and μ ;
do
  do
    Solve problem **P4** with the standard Lagrangian dual method ;
    Update $\boldsymbol{\lambda}, \mu$ as in (2.21) ;
  while ($\boldsymbol{\lambda}, \mu$ *do not converge*);
  Update $\tilde{\mathbf{y}}$ and Q_0 as in (2.27) ;
while ($\tilde{\mathbf{y}}$ *and* Q_0 *do not converge*);
for $\kappa = 1 : J$ **do**
  Find the first κ SBS's with largest values of $\tilde{y}_j$;
  Set $y_j = 1$ for these SBS's ;
  Calculate the EE ;
end
Select the κ that achieves the largest value of EE ;
Set $y_j = 1$ for the corresponding κ SBS's.

be rare; due to the constraint $x_{k,j} \leq y_j$, the number of y_j's in $(0, 1)$ would be small. Since the SBS's can be regarded as independent to each other, turning on the SBS's with the largest values of $\tilde{y}_j$ would achieve the highest EE. In particular, when the powers of all SBS's are the same, the proposed approach is optimal, since the set of SBS's that provide most performance gain are turned on. Based on Theorem 1, an upper bound for the complexity of Algorithm 1 is $J/(\varepsilon_1^2\varepsilon_2^2\varepsilon_3^2)$.

2.3.2 Optimal User Association with Given BS ON-OFF States and SBS Transmit Power

Considering that the timescale for updating of user association is much smaller than that of BS ON-OFF switching, user association is performed with a given set of BS ON-OFF states. With given SBS transmit power, the user association problem is formulated as

$$\mathbf{P6}: \max_{\{\mathbf{x}\}} \sum_{k=1}^{K} \sum_{j=0}^{J} x_{k,j} C_{k,j} \tag{2.32}$$

$$\text{s.t.: } (2.6) - (2.10)$$

Since $\{x_{k,j}\}$ and $\{y_j\}$ are 0-1 integers, a special property can be used to simplify the problem. Consider the constraint $x_{k,j} \leq y_j$. When $y_j = 1$, $x_{k,j} \leq y_j$ is always satisfied, and this constraint can be removed; when $y_j = 0$, $x_{k,j}$ must be 0 for all k. Thus, the SBS's that are turned off, i.e., $y_j = 0$, can be excluded from the problem formulation. Define Θ as the set of active SBS's, $\Theta = \{j \,|y_j = 1\}$. We re-index the active SBS's by $\{j = 1, ..., |\Theta|\}$. Same as **P2**,

we relax the integer constraints on $x_{k,j}$ and introduce the auxiliary variable $Q_0 = \sum_{k=1}^{K} x_{k,0}$. Problem **P6** can be reformulated as

$$\mathbf{P7}: \max_{\{\mathbf{x}\}} \left\{ \sum_{k=1}^{K} \sum_{j=1}^{|\Theta|} x_{k,j} \frac{R_{k,j}}{\sum_{k=1}^{K} x_{k,j}} + \sum_{k=1}^{K} x_{k,0} R_{k,0} \frac{T_u}{T'} - \sum_{k=1}^{K} x_{k,0} R_{k,0} Q_0 \frac{T_u}{T} \right\} \tag{2.33}$$

$$\text{s.t.: } \sum_{j=0}^{|\Theta|} x_{k,j} \leq 1, \ k = 1, 2, \ldots, K \tag{2.34}$$

$$\sum_{k=1}^{K} x_{k,j} \leq S_j, \ j = 0, 1, \ldots, |\Theta| \tag{2.35}$$

$$\sum_{k=1}^{K} x_{k,0} = Q_0 \tag{2.36}$$

$$0 \leq x_{k,j} \leq 1, \ k = 1, \ldots, K, \ j = 0, \ldots, |\Theta|. \tag{2.37}$$

In the objective function (2.33), the first term is non-convex. Based on the mobility of users, we consider the following two approaches to solve problem **P7**.

Low Mobility

In this case, we can use the value of $Q_j = \sum_{k=1}^{K} x_{k,j}$ in the previous period as an accurate approximation to the Q_j is the current period. Then, **P7** becomes a convex problem. We next show that the solution variables of problem **P7** are actually integers, although with the relaxed constraint (2.37).

As in problem **P2**, we use the Lagrangian dual method by taking a partial relaxation on the constraint $\sum_{k=1}^{K} x_{k,0} = Q_0$. Then, the optimal value of Q_0 can be obtained with (2.27). With a given Q_0, problem **P7** is transformed to an LP, denoted as problem **P8**. We then apply the same procedure as in Algorithm 1 to solve for $\mathbf{x}$.

Lemma 5 *All the decision variables in the optimal solution to the LP, problem* **P8***, are integers in* $\{0, 1\}$

The proof is omitted, as it is similar to that in [13, 44, 45].

High Mobility

We first introduce auxiliary variables Q_j, $j = 1, 2, ..., J$ and add $\sum_{k=1}^{K} x_{k,j} = Q_j$ as constraints. Then, we take partial relaxations on the constraints $\sum_{k=1}^{K} x_{k,0} = Q_j$, $j = 0, 1, ..., J$. Then, the local optimal Q_j for $j = 1, 2, ..., J$ can be obtained with the same subgradient approach in (2.27). With given Q_j, $j = 0, ..., J$, we solve the LP **P8** and obtain the suboptimal solution of **P7**.

2.3.3 Power Control with Given SBS ON-OFF States and User Association

The interference between different small cells is a major factor that impacts the EE, especially when the SBS's are densely deployed. To mitigate such interference, we employ a power control approach called iterative water-filling (IWF) [43]. As the multi-cell power control problem is non-convex, the IWF method uses the first-order derivative condition to derive the relations of powers of different BS's. The transmit power of SBS j on channel n, $P_j^T(n)$, is given as

$$P_j^T(n) = \left(\frac{1}{\nu_j + \psi_j(n)} - \frac{I_j(n) + \sigma^2}{\bar{H}_j(n)} \right)^+, \tag{2.38}$$

where ν_j is the Lagrangian multiplier corresponding to the constraint $P_j^T \leq P_{\max}^T$, $\psi_j(n)$ summarizes the effect of interference caused by SBS j to users in other SBS's, $I_j(n)$ accounts for the interference from other SBS's, $\bar{H}_j(n)$ is the channel power gain between SBS j and the user that uses channel n. In each small cell, the channels are randomly allocated to users.

We begin the iteration between $\{\mathbf{x}, \mathbf{y}\}$ and $\mathbf{P}_T$ with the case that all SBS's are turned on, in which the interference level is maximized. With the initial $\mathbf{y}$, we then obtain the initial $\mathbf{x}$ and $\mathbf{P}_T$. In the next iteration, we use the initial $\mathbf{P}_T$ to obtain $\{\mathbf{x}, \mathbf{y}\}$ by considering the SBS's that are still active. Thus, as the iterative process continues, more SBS's are turned off. As we can see from (2.38), $P_j^T(n)$ increases as the interference level decreases, and vice versa. Thus, $P_j^T(n)$ increases as more SBS's are turned off. With constraint $P_j^T \leq P_{\max}^T$, $P_j^T(n)$ is bounded for all n. Thus, the iteration process is guaranteed to converge.

2.4 Distributed Solutions

In a practical system, centralized control with global network information may not always be feasible due to constraints on complexity, overhead, or scalability. In this section, we propose two distributed schemes based on a user bidding approach and a wireless service provider (WSP) pricing approach, respectively.

2.4.1 User Bidding Approach

We assume that the utility of each user k is positively correlated to the achievable rate $C_{k,j}$ and user k always seeks to maximize $C_{k,j}$. The *preference list* of user k is determined by the $C_{k,j}$ values for different j. For instance, if $j^* = \arg\max_j\{C_{k,j}\}$, BS j^* is on top of user k's preference list. The preference list of BS j is also determined by $C_{k,j}$ in a similar way. Denote the price paid by user k to BS j as $p_{k,j}$. It is reasonable to assume that $p_{k,j}$ is an increasing function of $C_{k,j}$. The utility of BS j is defined as the payments made by all its connected users subtract the cost of power consumption q_j, given by

$$\sum_{k=1}^{K} x_{k,j}p_{k,j} - q_j. \tag{2.39}$$

To maximize the total utility under the constraint $\sum_{k=1}^{K} x_{k,j} = S_j$, each BS keeps the top S_j bids in its waiting list and reject the others. The repeated bidding game has two stages. In the *first* stage, each user bids for the top BS in its preference list. Receiving the bids, the BS's decide whether to hold or reject and bids and feedback the decision to users.

In the *second* stage, if a user has been rejected, the BS that rejected it would be deleted from its preference list. Then, the user bids for the most desirable BS among the remaining ones. Upon receiving the bids, each BS compares the new bids with those in its waiting list, and makes decisions on holding or rejecting the new bids. The rejected users then make another round of bids following the order of their preference lists, and the BS's again make decisions and feedback to users, and so forth. The bidding procedure is continued until convergence is achieved, i.e., the users in the waiting list of each BS do not change anymore. An upper bound

for the complexity of the bidding process is $J \cdot K$, which corresponds to the case that every user bids to every BS.

After convergence of the user association result, each SBS determines the value of its ON-OFF decision variable by comparing the payments and energy cost as follows.

$$y_j = \begin{cases} 1, & \text{if } \sum_{k=1}^{K} x_{k,j} p_{k,j} > q_j \\ 0, & \text{otherwise,} \end{cases} \quad j = 1, 2, \ldots, J. \tag{2.40}$$

It can be seen from (2.40) that SBS j chooses to be turned on only when it is profitable to do so. If SBS j is turned off, the users in the waiting list of SBS j will propose to the MBS. If the number of users in the waiting list of MBS exceeds S_0, the MBS would serve the top S_0 users with the largest SINRs.

It is obvious that the value of q_j impacts the system performance. When q_j is large, only the SBS's with a sufficient number of users to be served would be turned on due to the high energy cost; when q_j is small, more SBS's would be turned on, which potentially result in a low EE. We assume that q_j is predetermined using a database to find the value that maximizes the EE for a given relation between $p_{k,j}$ and $C_{k,j}$ and the traffic pattern.

Lemma 6 *The sequence of bids made by a user is non-increasing in its preference list.*

Proof: To maximize utility, a user first bids for the most desirable BS in its preference list. If rejected, the user deletes the BS from its preference list and bids for the most desirable BS in the updated list. Thus, the BS's chosen by a user is non-increasing in its preference list. ■

Lemma 7 *The sequence of bids in the waiting list of a BS is non-decreasing in its preference list.*

Proof: According to the strategy of each BS, if a BS is not fully loaded, it put all the bids into the waiting list. If a BS is fully loaded, it compares the new incoming bids with the bids already in the list, and selects the most profitable bids to maximize its own utility. ■

Theorem 2 *The repeated game converges.*

Proof: Suppose the game does not converge. Then, there must be a user k and a BS j such that: (i) user k prefers BS j to its current connecting BS j', (ii) BS j prefers user k to user k', who is currently in the waiting list of BS j. Under this circumstance, user k is a better choice and BS j can accept the bid of user k. User k will bid for BS j.

Based on Lemma 7, the sequence of bids received by BS j is non-decreasing. As user k is a better choice than user k' for BS j while user k is not in the waiting list, it must be the case that user k has never bidden for BS j. Since user k prefers BS j to BS j', user k must bid for BS j prior to BS j'. We conclude that user k also has never bidden for BS j'. However, user k is currently in the waiting list of BS j', indicating that user k has bidden for BS j' before, which is a contradiction. Thus, we conclude that the repeated game converges. ■

Lemma 8 *During any round of the repeated game, if user k bids for BS j, it cannot have a better choice than BS j.*

Proof: According to the bidding strategy of users, if user k bids for BS j, either BS j is its most preferable choice or user k has been rejected by another BS j'. The reason BS j' rejects user k is because it already held S_j bids that are better than user k. Since the sequence of bids for each BS is non-decreasing, it is impossible for user k to enter the waiting list of BS j'. Thus, user k can not have a better choice than BS j. ■

Theorem 3 *The repeated game converges to an NE that is optimal for each user and BS.*

Proof: Based on the strategy of BS's, each BS holds the set of users with the maximum sum payments. For an SBS, if the sum of user payments is less than its power cost, the optimal strategy is to sleep so that the utility is increased from a negative value to zero.

From Lemma 8, if a user is currently in the waiting list of a BS, this BS is the best possible option for the user. Thus, when the game converges, the outcome is the best response of each user. Following Theorem 2, we conclude that the repeated bidding game converges to an NE. ■

2.4.2 Service Provider Pricing Approach

Although the proposed user bidding based approach can be implemented by each user and SBS in a distributed manner, the bidding process generates frequent information exchange between users and SBS's. To avoid such overhead, we propose a WSP pricing approach in this part by formulating a game between users and WSP. In the pricing game, the WSP sets the price of each BS for each user, with the objective of maximize its utility. Then, each user decides which BS to connect to based on the achievable rate and price. Finally, the SBS's determine their ON-OFF states by comparing the total payments with energy cost. Compared to the user bidding approach, the WSP pricing approach has a lower communication overhead, but requires more computation at the MBS.

Since the MBS is always turned on to guarantee the basic communication requirements of users, we assume all users pay a pre-determined, constant price p_0 for connecting to MBS, i.e., $p_{k,0} = p_0$, for all k. Based on p_0, user k pays an additional fee of $\eta_{k,j}$ for connecting to SBS j, with the expectation of achieving a higher rate. Thus, if user k choose to connect to SBS j, the total price would be $p_{k,j} = p_0 + \eta_{k,j}, j = 1, 2, ..., J$. Let the satisfaction level of a user be a logarithmic function of the achievable rate to capture the diminishing marginal effect [31], then the utility of user k is

$$\mathcal{U}_k = \sum_{j=0}^{J} x_{k,j} \left\{ w_k \log \left(C_{k,j} \right) - p_{k,j} \right\}, \tag{2.41}$$

where w_k is a weight that interprets user's satisfaction level to monetary utility.

Due to the constraint $\sum_{j=0}^{J} x_{k,j} \leq 1$, the strategy of a user is to choose the BS that provides the maximum utility. Compared to connecting to the MBS, the additional utility of user k obtained by connecting to SBS j is $w_k \log \left(C_{k,j} \right) - w_k \log \left(C_{k,0} \right) - \eta_{k,j}$.

Denote the SBS that provides the maximal utility to user k as j^*, which can be expressed as

$$j^* = \underset{\{j=1,...,J\}}{\arg\max} \left\{ w_k \log \left(C_{k,j} \right) - \eta_{k,j} \right\}. \tag{2.42}$$

Thus, the strategy of user k is given as

$$\begin{cases} x_{k,j^*} = 1, \text{ if } w_k \log\left(C_{k,j}\right) - w_k \log\left(C_{k,0}\right) \geq \eta_{k,j} \\ x_{k,0} = 1, \text{ otherwise.} \end{cases} \tag{2.43}$$

Here, we assume that a user chooses the SBS when the achievable utility is equal to that of the MBS. From (2.43), it can be easily verified that the highest payment obtained by SBS j from user k is $w_k \log\left(C_{k,j}\right) - w_k \log\left(C_{k,0}\right)$.

Define the utility of WSP as the total payments obtained from users subtract the cost of BS power consumption, $\mathcal{U}_{WSP} = \sum_{k=1}^{K} \sum_{j=0}^{J} x_{k,j} p_{k,j} - \sum_{j=1}^{J} y_j q_j$. We assume q_j and w_k are the same for all j and k, respectively. Then, the performance is directly determined by w_k/q_j. We also assume that the optimal w_k/q_j that achieves the highest EE is predetermined using a database. Since the MBS is always turned on, and each user pays a fixed amount for MBS connection, the utility maximization of WSP is equivalent to the following problem.

$$\mathbf{P9}: \max_{\{\boldsymbol{\eta},\mathbf{x},\mathbf{y}\}} \sum_{k=1}^{K} \sum_{j=1}^{J} x_{k,j} \eta_{k,j} - \sum_{j=1}^{J} y_j q_j \tag{2.44}$$

$$\text{s.t.: } (2.6) - (2.11) \text{ and}$$

$$\eta_{k,j} \leq w_k \log\left(C_{k,j}\right) - w_k \log\left(C_{k,0}\right), \quad k = 1, 2, \ldots, K, \ j = 1, 2, \ldots, J. \tag{2.45}$$

Problem **P9** is difficult to solve directly since $\mathbf{x}$ is coupled with both $\boldsymbol{\eta}$ and $\mathbf{y}$. However, using the property of the user association strategy given in (2.43), it is possible for decouple $\mathbf{x}$ and $\boldsymbol{\eta}$ through pricing strategy.

Lemma 9 *The user association can be controlled by WSP with the following pricing strategy.*

$$\eta_{k,j} = \begin{cases} \Delta_{k,j}, & \text{if } x_{k,j} = 1 \\ \Delta_{k,j} + \varepsilon, & \text{otherwise} \end{cases}, \tag{2.46}$$

where $\Delta_{k,j} = w_k \log\left(C_{k,j}\right) - w_k \log\left(C_{k,0}\right)$*,* ε *is an arbitrary positive number.*

Proof: For a user-SBS pair (k, j) desired by the WSP, the price $\eta_{k,j}$ is set to the additional utility achieved by the increased data rate, $\Delta_{k,j}$. The additional utility that can be achieved by the user is 0. As in (2.43), this user-SBS pair would be associated. For a user-SBS pair (k, j) not selected, the WSP sets the price $\eta_{k,j}$ to a value larger than $\Delta_{k,j}$. Hence, user k would not connect to SBS j since less utility can be obtained than connecting to either the MBS or another SBS. ■

Denote the objective value of problem **P9** as $\mathcal{U}'_{WSP}$, since $\eta_{k,j} \leq \Delta_{k,j}$ holds for all k and j, an upper bound of $\mathcal{U}'_{WSP}$ is given as $\mathcal{U}' = \sum_{k=1}^{K} \sum_{j=1}^{J} x_{k,j}\Delta_{k,j} - \sum_{j=1}^{J} y_j q_j$.

Lemma 10 *The upper bound of $\mathcal{U}'_{WSP}$ is achievable if the WSP adopts the pricing strategy given in (2.46).*

Proof: With the pricing strategy described in (2.46), the equality $\Delta_{k,j} = \eta_{k,j}$ holds for all the (k, j) pairs with $x_{k,j} = 1$. Thus, with $\mathbf{x}$ and $\mathbf{y}$ as variables and other constraints remaining the same, the maximum value of $\mathcal{U}'_{WSP}$ equals to the maximum value of $\mathcal{U}'$. ■

From Lemma 10, we can see that maximizing $\mathcal{U}'_{WSP}$ is equivalent to maximizing $\mathcal{U}'$ with the pricing strategy given in (2.46). Problem **P9** is reduced to the following problem.

$$\mathbf{P10}: \max_{\{\mathbf{x},\mathbf{y}\}} \sum_{k=1}^{K} \sum_{j=1}^{J} x_{k,j}\Delta_{k,j} - \sum_{j=1}^{J} y_j q_j \tag{2.47}$$

$$\text{s.t.: } (2.6) - (2.11).$$

Since $\Delta_{k,j}$ is coupled with $x_{k,j}$, we use an iterative approach to decouple these two variables with proven optimality. Let $\mathbf{C}_0 = [C_{1,0}, C_{2,0}, ..., C_{K,0}]^T$. At each iteration, we solve problem **P10** for given values of $\mathbf{C}_0$. Then, $\mathbf{C}_0$ is updated after each iteration until convergence. For fixed values of $\log(C_{k,0})$, $\Delta_{k,j}$ are also fixed for all k and j. Problem **P10** has a similar structure with the ones presented in Section 2.3. Therefore, we can apply the same decomposition approach to obtain the optimal solution for $\mathbf{x}$ and $\mathbf{y}$. Given the solution of $\mathbf{x}$, the optimal pricing strategy for WSP can be determined by (2.46). The iterative approach is described in Algorithm 2.

The idea of Algorithm 2 is to search different values of $\mathbf{C}_0$ and obtain the corresponding $\mathbf{x}$. The search terminates until $\mathbf{x}$ matches the expressions of $\mathbf{C}_0$. At the beginning, we set $C_{k,0}^{[0]} = \max\left\{\left(1 - \frac{KT'}{T}\right)\frac{T_u}{T'}\log\left(1+\gamma_{k,0}\right), 0\right\}$, which corresponds to the case that all users are connected to MBS. Given the initial $\mathbf{C}_0$, we solve problem **P10** and select a certain set of users to be served by SBS's based on the solution. As the initial $C_{k,0}$ is set to the lowest possible value for each k, such a solution achieves the maximum value of $\sum_{k=1}^{K}\sum_{j=1}^{J} x_{k,j}\Delta_{k,j}$. This is because a maximum number of users are selected to connect to SBS so that each $\Delta_{k,j}$ becomes its largest possible value. Then, for each user, $C_{k,0}$ is updated to a higher value in the first iteration with (2.3). After updating $\{C_{k,0}\}$, we solve problem **P10** in the second iteration. With a higher value of $C_{k,0}$, $\Delta_{k,j}$ is decreased for all k and j and some the user-SBS pairs would have negative $\Delta_{k,j}$. Then, these user-SBS pairs would not be selected by WSP due to their negative utilities. As a result, these users would switch to the MBS, and the updated values of $\{C_{k,0}\}$ would be decreased compared to the ones in the first iteration. In the next iteration, some users would be selected to connect to the SBS's again due to the decreased values of $\{C_{k,0}\}$, which in turn increases the values of $\{C_{k,0}\}$ since fewer users are connected to the MBS. Such a process is repeated with all $C_{k,0}$ increase and decrease alternatively until $\mathbf{C}_0$ converges.

Lemma 11 *Algorithm 2 converges to a solution with (2.3) holds for all k.*

Proof: Case 1 (Special Case): With the initial $\mathbf{C}_0$, the solution of problem **P10** already satisfies the relation between $\mathbf{C}_0$ and $\mathbf{x}$ given by (2.3). Since there is only one iteration, it is obvious that Lemma 11 holds under this case.

Case 2 (General Case): Suppose more than one iterations are required, then $\mathbf{C}_0$ would be updated for more than one times. As the initial values of $\{C_{k,0}\}$ are set to the lowest, we have $\mathbf{C}_0^{[0]} \prec \mathbf{C}_0^{[t]}$ for $t \geq 1$. In particular, $\mathbf{C}_0^{[0]} \prec \mathbf{C}_0^{[2]}$. This indicates that some users that are connected to SBS's in the first iteration would not switch to the MBS in the second iteration. Thus, the number of users that switch between SBS and MBS is decreased from the first to the second iteration. Regarding $\mathbf{C}_0^{[2]}$ as a set of initial values and applying the same analysis, we have $\mathbf{C}_0^{[0]} \prec \mathbf{C}_0^{[2]} \prec \mathbf{C}_0^{[4]} \prec \mathbf{C}_0^{[6]} \prec \cdots$. The same result holds for the case when t is an odd number using a similar analysis. Thus, the number of users that switch between SBS and MBS

is decreasing for $t \geq 1$, and the number would become zero after a finite number of iterations. This means the solution of $\mathbf{x}$ will converge. ■

Lemma 12 *Suppose* $\mathbf{C}_0$ *converges after the* t^**th iteration. Then,* $\mathbf{C}_0^{[t^*]}$ *is the unique vector that satisfies* (2.3) *for all* k.

Proof: Suppose there is another $\mathbf{C}_0^{[t']}$ that satisfies (2.3). Without loss of generality, we assume $\mathbf{C}_0^{[t']} \succ \mathbf{C}_0^{[t^*]}$. On one hand, with (2.3), we have $\sum_{k=1}^{K} x_{k,0}^{[t']} < \sum_{k=1}^{K} x_{k,0}^{[t^*]}$. Then, we have $\sum_{k=1}^{K}\sum_{j=1}^{J} x_{k,j}^{[t']} > \sum_{k=1}^{K}\sum_{j=1}^{J} x_{k,j}^{[t^*]}$. On the other hand, since $\mathbf{C}_0^{[t']} \succ \mathbf{C}_0^{[t^*]}$, we have $\Delta_{k,j}^{[t']} < \Delta_{k,j}^{[t^*]}$, $k = 1, 2, ..., K$, $j = 1, 2, ..., J$. As a result, the number of user-SBS pairs with $\Delta_{k,j}^{[t']} < 0$ is no less than the number of user-SBS pairs with $\Delta_{k,j}^{[t^*]} < 0$. As discussed, the user-SBS pairs with negative $\Delta_{k,j}$ would not be selected to connect to the SBS due to their negative utilities. Thus, we have $\sum_{k=1}^{K}\sum_{j=1}^{J} x_{k,j}^{[t']} < \sum_{k=1}^{K}\sum_{j=1}^{J} x_{k,j}^{[t^*]}$, a contradiction.

For the case $\mathbf{C}_0^{[t']} \prec \mathbf{C}_0^{[t^*]}$, we will also get a contradiction. Combine these two cases, we conclude that $\mathbf{C}_0^{[t^*]}$ is the only feasible vector. ■

Theorem 4 *Algorithm 2 achieves the optimal solution to problem* **P10**.

Proof: According to Lemma 12, when $\mathbf{C}_0$ converge, the relation between $\mathbf{C}_0$ and $\mathbf{x}$ described in (2.3) holds for all k. Thus, we can solve problem **P10** by fixing $\mathbf{C}_0$. Since such $\mathbf{C}_0$ is unique, the optimal solution of problem **P10** can be obtained with the procedure in Algorithm 2. Since both user and the WSP achieve the maximum utility, we conclude the proposed pricing game achieves an NE. ■

In Algorithm 2, each user determines which BS to connect in a distributed way. However, the SBS ON-OFF decision is still made by WSP with a centralized approach. To enable each SBS to make its own decision in a distributed pattern, we propose a modified pricing-based user association and SBS ON-OFF strategy. In the modified pricing scheme, we adopt the same iterative approach as the original pricing scheme to decouple $\mathbf{x}$ and $\mathbf{C}_0$ so that the process is guaranteed to converge. Different from the original pricing scheme, the user association is determined by solving the following problem.

Algorithm 2: WSP Pricing based User Association and SBS ON-OFF Strategy

Initialize $t = 0$;
for $k = 1 : K$ **do**
  $C_{k,0}^{[0]} = \max\left\{\left(1 - \frac{KT'}{T}\right)\frac{T_u}{T'}\log(1+\gamma_{k,0}), 0\right\}$;
end
Obtain $C_{k,j}$, $k = 1, 2, ..., K$, $j = 1, 2, ..., J$ from SBS's ;
do
  for $k = 1 : K$ **do**
    for $j = 1 : J$ **do**
      $\Delta_{k,j}^{[t+1]} = w_k \log\left(C_{k,j}\right) - w_k \log\left(C_{k,0}^{[t]}\right)$;
    end
  end
  Obtain the optimal $\mathbf{x}^{[t+1]}$ and $\mathbf{y}^{[t+1]}$ by solving problem **P10** ;
  for $k = 1 : K$ **do**
    Update $\mathbf{C}_0$ as $C_{k,0}^{[t+1]} = \left(1 - \sum_{k=1}^{K} x_{k,0}^{[t+1]}\left(\frac{T'}{T}\right)\right)\frac{T_u}{T'}\log\left(1+\gamma_{k,0}\right)$;
  end
  $t = t + 1$;
while ($\mathbf{x}$ *does not converge*);
for $k = 1 : K$ **do**
  for $j = 1 : J$ **do**
    WSP sets $\Delta_{k,j}$ according to (2.46) ;
  end
end
for $k = 1 : K$ **do**
  Each user determines which BS to connect to according to (2.43) ;
end

$$\mathbf{P11}: \max_{\{\mathbf{x}\}} \sum_{k=1}^{K}\sum_{j=1}^{J} x_{k,j}\Delta_{k,j}, \quad \text{s.t.: } (2.6) - (2.10). \tag{2.48}$$

As discussed in Section 2.3, the constraint matrix of problem **P11** is unimodular, the optimal solution of **P11** can be obtained by relaxing the integer constraint and solving the linear programming problem. With the solution of $\mathbf{x}$, each SBS determines its ON-OFF state with the same strategy as we presented in the bidding game, which is given in (2.30).

Compared to problem **P10**, the objective of problem **P11** is to maximize the total payments received by WSP, which does not account for the cost of BS power consumption. Thus, the modified pricing scheme does not achieve an NE for the users and WSP since the WSP may not achieve the maximum utility. However, we will show in simulations that the performances of the modified pricing scheme are close to that of the original pricing scheme.

2.5 Simulation Study

We evaluate the proposed centralized and distributed schemes with MATLAB simulations. We use the path loss and SINR models in [12]. The path loss is $(1 + \left(\frac{d}{40}\right)^{3.5})^{-1}$ between MBS and a user and $(1 + \left(\frac{d}{40}\right)^{4})^{-1}$ between an SBS and a user, and the channel experience Rayleigh fading with unit mean power [12]. A 1000m $\times$ 1000m area is used. The massive MIMO BS is located at the center, the SBS's are randomly distributed in the area. We consider two cases for user distribution. In the first case, users are uniformly distributed across the area. In the second case, we divide the area into 8 subareas, the number of users in each subarea is a Poisson random variable and the users in each subarea are randomly distributed. Then, we have different user densities in these subareas. The maximum powers of the MBS and SBS's are set to 40 dBm and 30 dBm, respectively. The number of channels is 50 for SBS's, thus $S_j = 50$ for $j = 1, 2, \ldots, J$. We also set $S_0 = 100$.

We compare with two heuristic schemes for BS ON-OFF switching strategy. *Heuristic 1* is based on a load-aware strategic BS sleeping mode proposed in [49]. Specifically, SBS j is turned on with probability $\min\{\theta_j/S_j, 1\}$, where θ_j is the number of users within the coverage of SBS j. *Heuristic 2* is based on a scheme presented in [50], where an SBS is activated whenever there is a user enters its coverage area. We also consider the case that all the SBS's are always active as a benchmark (termed Always ON). For the Always ON and two heuristic schemes, the user association strategy is determined by the solution in Section 2.3-B. For the distributed schemes, the parameters w_k and q_j are set to be the optimal values that achieve the highest EE.

The EEs of different schemes are presented in Figs. 2.1–2.4. In Figs. 2.1 and 2.2, it can be seen that the EEs of Always ON and Heuristic 2 schemes decrease when the number of SBS's becomes large, due to the fact that some SBS's become under-utilized. The EEs of the proposed schemes and Heuristic 1 do not decrease as the number of SBS's grows, since these schemes can dynamically adjust to the traffic demand and turn off the under-utilized SBS's. As expected, the centralized scheme achieves the highest EE. Note that the EE of Heuristic 2 is close to the Always ON scheme, since an SBS is easily activated when the numbers of users and SBS's are

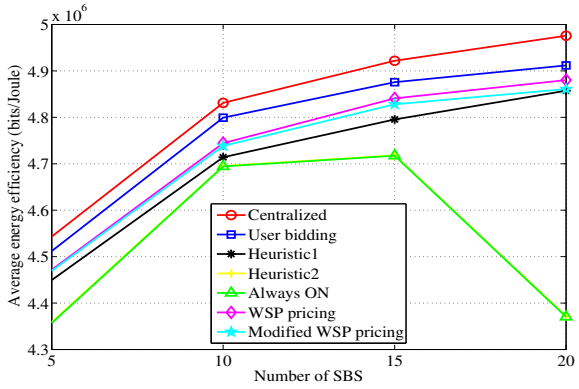

Figure 2.1: Average system EE versus number of SBS's for different BS ON-OFF switching strategies: 100 users, uniformly distributed.

sufficiently large. The two distributed schemes also achieve high EEs, since activation of SBS's depends on the payments received from users connecting to these SBS's. The EE of the user bidding scheme is slightly higher than that of the pricing scheme since each user and SBS has a preference list, the propose and reject processes contribute to a better matching between users and SBS. For the pricing scheme, the decision of users are controlled by the MBS. It is more likely that a user located at the edge of different small cells are served by an SBS with high load while the SBS is not the optimal choice for the user. We also find that the performance of the modified pricing scheme is close to that of the original pricing scheme, especially when the number of SBS is small, showing that the decision made by each SBS is close to the centralized decision made by WSP. Compare Figs. 2.1 and. 2.2, it can be seen that when the traffic load is varying over subareas, the gaps between the proposed schemes and other schemes are slightly increased since larger gains can be achieved when the traffic demand becomes geographically dynamic.

Figs. 2.3 and 2.4 show the EE performance under different numbers of users. We also find that the proposed schemes outperform the other schemes, while the gaps become smaller as the number of users grows. This is because when the traffic load increases, more SBS's would be activated with the proposed scheme, since they can effectively offload the traffic load from MBS and significantly enhance the sum rate. In case with extremely large number of users, the Always ON scheme would be optimal.

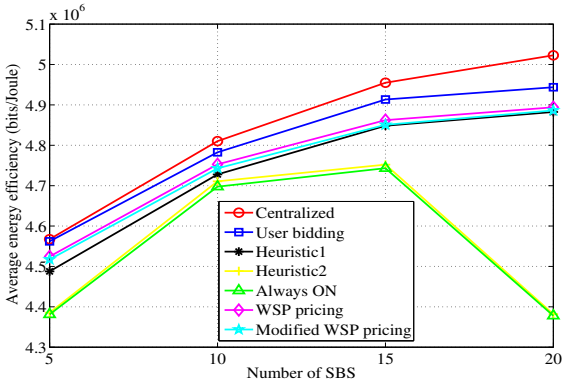

Figure 2.2: Average system EE versus number of SBS's for different BS ON-OFF switching strategies: 100 users, non-uniformly distributed.

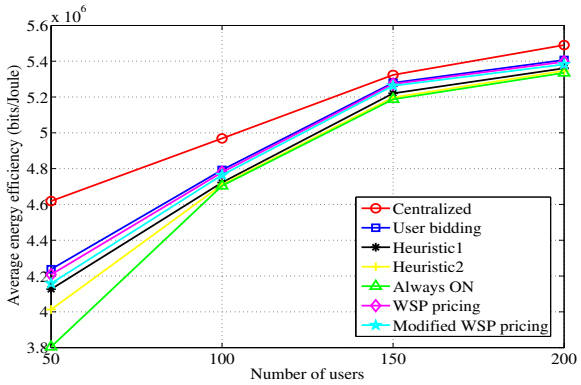

Figure 2.3: Average EE efficiency versus number of users for different BS ON-OFF switching strategies: uniformly distributed users, 10 SBS's.

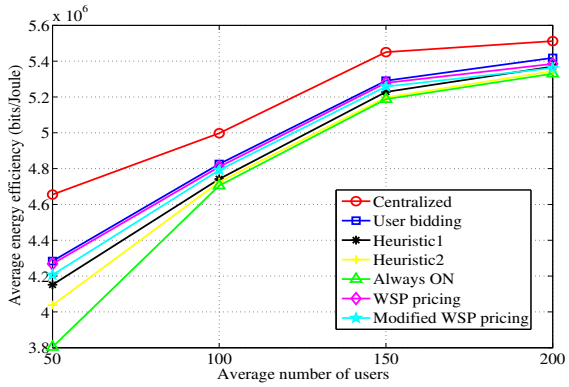

Figure 2.4: Average system EE versus average number of users for different BS ON-OFF switching strategies: non-uniformly distributed users, 10 SBS's.

We also evaluate the sum rate of all schemes in Fig. 2.5. We find that the sum rate is improved as more SBS's are deployed, due to more offloading and higher average SINR. Obviously, Always ON offers the best performance since it is possible for each user to connect to

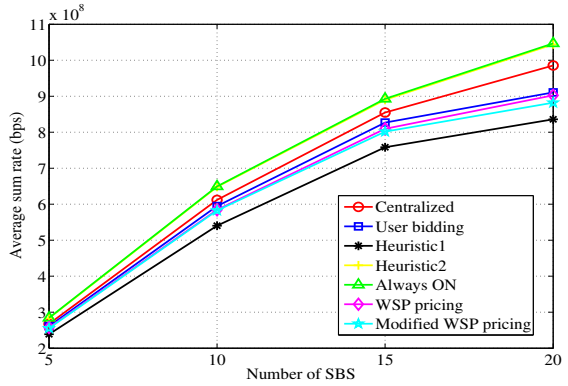

Figure 2.5: Average sum rate versus number of SBS's for different BS OF-OFF switching strategies: 100 users, uniformly distributed.

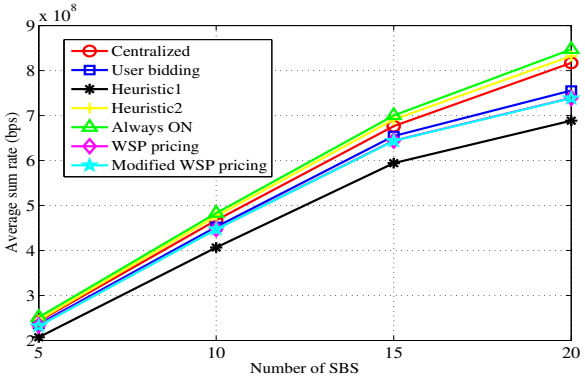

Figure 2.6: Average sum rate versus number of SBS's for different BS ON-OFF switching strategies: 100 users, non-uniformly distributed.

the BS with the largest achievable rate. The sum rate of the centralized scheme is close to that of Always ON. This is because we choose to turn off the SBS's that are not energy efficient, i.e., the sum rates of users connecting to these SBS's are not large enough and it is not worthy to turn on these SBS's. The two distributed schemes also achieve a high sum rate performance, because the SBS's with negative utility are turned off. Since the sum rates of SBS's that are turned off are relatively small, the performance loss is small.

In Fig. 2.9, an example of the repeated user bidding game is given. It can be seen that the game converges after a few number of rounds with the proposed algorithm. Note that, after the bidding game converges after 6 rounds, the utility of the BS's is slightly increased, due to the fact that some SBS's with negative utility are turned off. The utility of users is decreased since some SBS's are turned off and their users are handed over to the MBS.

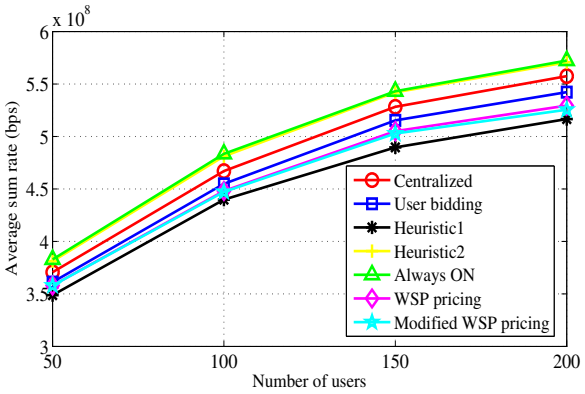

Figure 2.7: Average sum rate versus number of users for different BS ON-OFF switching strategies: uniformly distributed users, 10 SBS's.

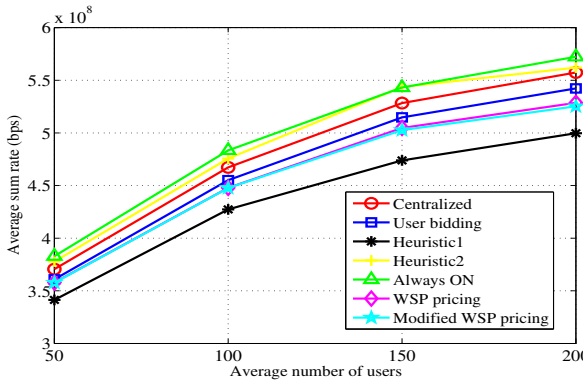

Figure 2.8: Average sum rate versus number of users for different BS ON-OFF switching strategies: non-uniformly distributed users, 10 SBS's.

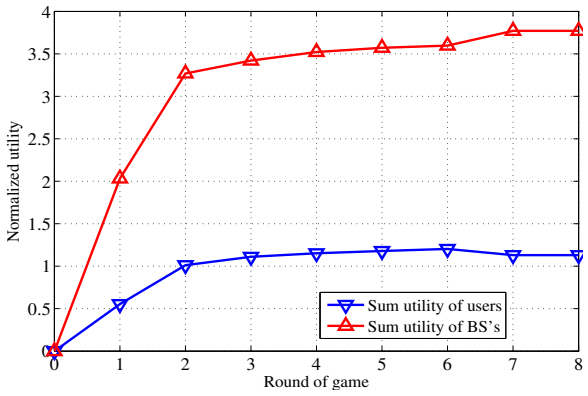

Figure 2.9: Convergence of the repeated bidding game: 100 users and 10 SBS's.

Fig. 2.10 shows the convergence of the proposed pricing scheme. The utility of WSP increases and decreases alternatively and finally converges to a unique value after several iterations.

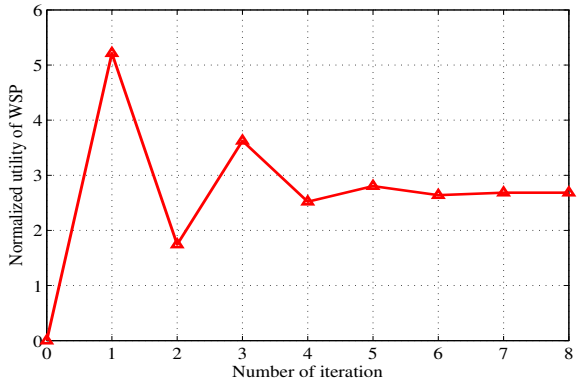

Figure 2.10: Convergence of the iterative pricing scheme: 100 users and 10 SBS's.

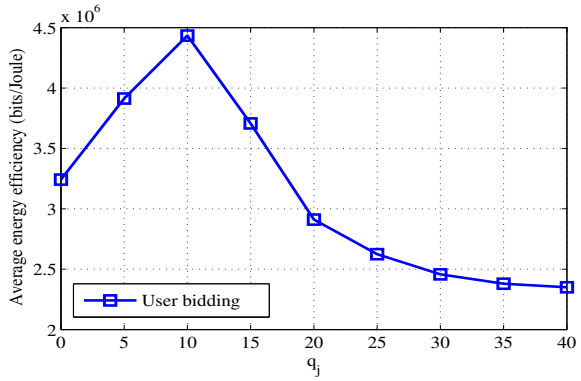

Figure 2.11: Average system EE versus different values of q_j : 100 users and 10 SBS's, $p_{k,j} = 1$.

The impact of q_j is shown in Fig. 2.11. When q_j is small, most SBS's would be turned ON since the cost of energy consumption is low for each SBS. As q_j increases, some SBS's would be turned OFF to save energy, resulting improved EE. When q_j is set to be around 10, maximum EE can be achieved, since the SBS's with relatively low sum rate are turned OFF. As q_j continues to increase, the system EE is declined, since the higher cost makes more SBS's to turn OFF, resulting decreased sum rates of users and EE.

2.6 Related Work

As key technologies for 5G network, massive MIMO and small cells have been extensively studied in prior works. The fundamental PHY layer techniques of massive MIMO were introduced in [8, 61]. Beyond the PHY, upper layer techniques in a wireless network with massive MIMO are also considered, such as [3, 13, 28]. In [13], user association and resource allocation

in a massive MIMO HetNet were investigated with the objectives of rate maximization and rate maximization with proportional fairness. In [28], a time-shift frame structure was proposed to mitigate inter-cell interference caused by pilot contamination in a multi-cell massive MIMO system. Since neighboring cells transmit pilots at different time instants, the inter-cell interference can be well mitigated.

User association in HetNet has been widely investigated, such as [31, 77, 81]. In [31], user association and resource allocation were jointly considered to maximize the sum utility of users. Using dual decomposition, the proposed scheme can be implemented with a distributed algorithm, and the solution is shown to be near-optimal. In [81], user association was considered to minimize the maximum load among all BS's, several approximation algorithms were proposed with analysis on complexity and performance bound. In [77], user association is determined by the achievable rate of each user. The HetNet with dense small cell deployment has drawn increasing interests, especially for ultra-dense HetNet. Several cooperative approaches were proposed in [48] and [54] to enhance the network. performance.

The EE has become an important objective for wireless networks in recent years. Specifically, the designs of energy-efficient massive MIMO systems were studied in [**?**, **?**], where power and subcarrier allocation, antenna selection, and pilot allocation were considered to maximize EE. Some prior works also aimed to improve the EE of HetNets [49]. In [49], the authors considered two sleeping strategies for MBS, and derived the success probability and EE for a K-tier heterogeneous network using stochastic geometry analysis. The BS ON-OFF switching strategie was also investigated by prior works including [37]. In [37], a distributed algorithm that is easy to implement was proposed. The key principle is the use of a new notion called *network-impact*, which accounts for the load increments brought to other BS's by turning a BS OFF. Compared to these works, we focus on the optimal BS ON-OFF switching strategy for energy efficient massive MIMO enabled HetNets.

2.7 Conclusions

In this chapter, we considered BS ON-OFF switching, user association, and power control to maximize the EE of a massive MIMO HetNet. We formulated an integer programming problem

and proposed a centralized scheme to solve it with near optimal solution. We also proposed two distributed schemes based on a user bidding approach and a WSP pricing approach. We showed that an NE can be achieved for these two distributed schemes. The proposed schemes were evaluated with simulations and the results demonstrated their superior performance over benchmark schemes.

Chapter 3

Interference Management and User Association in a Massive MIMO HetNet: A Nested Array Approach

3.1 Introduction

Massive MIMO (Multiple Input Multiple Output) and *small cell* are recognized as two key technologies for 5G wireless systems due to their great potential to enhance network capacity [2]. In a massive MIMO system, the base station (BS) is equipped with more than 100 antennas and serves multiple users with the same spectrum band [8]. With aggressive spatial multiplexing, a massive MIMO can dramatically improve both energy and spectral efficiency compared to traditional wireless systems [3, 9, 10, 61]. On the other hand, small cell deployment achieves high signal to noise ratio (SNR) and high spectrum spatial reuse due to short transmission distance and small coverage area. As a result, a heterogeneous network (HetNet) with small cells can significantly boost network capacity compared to traditional macrocell network.

Due to these benefits, massive MIMO HetNet, which integrates these two techniques, has drawn considerable attention recently [**?**, 4, 11–13, 24, 27, 52, 87]. A massive MIMO HetNet consists of a macrocell BS (MBS) and multiple small cell BS's (SBS), the MBS is equipped with a large number of antennas. Due to spectrum scarcity in cellular networks, small cells are expected to share the same spectrum band with the macrocell, resulting in cross-tier interference. While interference management in a regular HetNet mainly focuses on resource allocation in the time-frequency domain [54], the spatial characteristics of massive MIMO can be exploited to mitigate interference in massive MIMO HetNets. In [6], a spatial blanking scheme was proposed in which the transmission energy of the MBS is focused on certain directions that do not cause interference to small cells. In [52], a reversed time division duplex (RTDD)

architecture was introduced. Since the channels between the MBS and SBS's are quasi-static, the MBS can carry out zero-forcing beamforming based on the estimated channel covariance. In [11], coordinated transmissions are assumed between the MBS and SBS's, so that each user receives signals from both the MBS and SBS's. Through coordinated beamforming of BS's, the interference between different transmissions can be minimized.

The interference management in most existing works are performed with baseband processing at the digital domain, which requires channel state information (CSI) between the interfering transceivers. However, in a massive MIMO HetNet with dense small cell deployment, acquiring the CSI of interfering links is difficult and causes large overhead due to the large number of antennas at MBS and the large number of SBS's. To overcome this drawback, an efficient approach is to mitigate inter-cell interference with antenna array processing at the analog domain and deal with intra-cell interference with baseband processing at the digital domain. With antenna array processing, the directions of arrival (DoA) of interference sources can be estimated. Then, a beamforming with respect to different directions can be applied to nullify the interference from certain directions. However, unlike typical application scenarios that employ antenna array for DoA estimation, e.g., radar system, the DoA estimation in wireless communication network is highly challenging due to the rich scattering and multipath effect. As the signals of a user received by an SBS comes from multiple directions, the antenna array needs to resolve large number of DoAs. However, with traditional antenna array configuration such as uniform linear array, the number of directions that can be resolved is $O(N)$, which is insufficient for a massive MIMO HetNet. To this end, we employ a second order antenna array processing technique called *nested array* [55] for inter-cell interference management in massive MIMO HetNets. Based on the concept of *difference co-array*, a nested array is implemented by nonuniform antenna placement to achieve $O(N^2)$ degrees of freedom (DoF) with only N antennas. Further, the DoA estimation is performed with a *passive sensing* pattern, i.e., the antenna array does not need to send out signals for detection. Due to the significantly increased DoF and its easy implementation, we apply nested array to identify a number of $O(N^2)$ directions of incoming signals, which include both desired signals and interference.

Then, the signals from different directions can be filtered such that the desired signals remain while interference signals are nullified.

Due to the multipath effect, the signals of a user received by an SBS comes from multiple directions. Then, a certain number of DoFs are required to estimate the DoAs of these signals. Hence, both the channel gain and the number of multipaths of each user should be taken into consideration to enhance the system performance. Since both service provisioning and interference nulling require the use of DoF, there is a tradeoff between these two objectives. When an SBS serves more nearby users, the sum rate would increase, however, the available DoF for interference nulling would be reduced, resulting in degraded signal to interference and noise ratio (SINR). Thus, given the DoFs of each SBS, a key design problem is to select the set of users to be served and the set of interfering users for interference nulling, to optimize the system performance. In this chapter, we consider user association and interference nulling scheduling in a nested array-based massive MIMO HetNet to fully harness the benefit of interference mitigation brought about by nested array. The main contributions of this chapter are summarized as follows.

- We formulate the user association and interference nulling problem as an integer programming problem with the objective of maximizing the sum rate of a massive MIMO HetNet, subject to constraints on the DoF of each BS.

- We first consider interference nulling schedule with *a given user association*. The resulting integer programming problem has a nonlinear and nonconvex objective function, we propose a series of approximations that transform the original problem into an integer programming problem with a linear objective function. The optimal solution to such a problem can be obtained with a *cutting plane* approach. Moreover, we find that in the high SINR regime and when each BS only receives a fixed number of strongest signals from each user, the constraint matrix will become *unimodular* and the integer programming problem will become equivalent to an linear programming (LP) problem obtained by relaxing the integer constraints. Thus an optimal solution can be obtained with an LP solver.

- We then consider *joint interference nulling schedule and user association*. To address the highly complicated structure of the original problem, we propose a distributed scheme based on a poly matching between users and BS's. We show that the matching process converges and the outcome yields a stable matching that is optimal for each user and BS.

- The proposed schemes are compared with other benchmarks schemes through simulations. The results show that near optimal performance can be achieved.

The remainder of this chapter is organized as follows. The nested array based interference nulling method is introduced in Section 3.2. The system model and problem formulation are presented in Section 3.3. The solution for interference nulling with a given user association is presented in Section 3.4. The distributed algorithm for joint interference nulling scheduling and user association is presented in Section 3.5. The simulation results are discussed in Section 3.6. We present related works in Section 3.7 and conclude this chapter in Section 3.8.

3.2 Preliminaries

3.2.1 Signal Model of Difference Co-Array

Consider an antenna array with N antennas, the $N \times 1$ steering vector corresponding to direction θ is denoted as $\mathbf{a}(\theta)$. Let d_i be the position of the ith antenna and λ the carrier wavelength. The ith element of $\mathbf{a}(\theta)$ is $e^{j(2\pi/\lambda)d_i \sin\theta}$. Suppose D narrowband sources from directions $\{\theta_i, i = 1, 2, ..., D\}$ impinge upon the antenna array with powers $\{\sigma_i^2, i = 1, 2, ..., D\}$. The received signal is given by

$$\mathbf{r}[m] = \mathbf{F}\boldsymbol{\gamma}[m] + \mathbf{n}\left[m\right], \ m = 1, 2, ..., N, \tag{3.1}$$

where $\boldsymbol{\gamma}[m]_{D\times 1} = \left[\gamma_1\left[m\right], \gamma_2\left[m\right], ..., \gamma_D\left[m\right]\right]^T$ is the source signal vector,
$\mathbf{F} = \left[\mathbf{f}\left(\theta_\mathbf{1}\right), \mathbf{f}\left(\theta_\mathbf{1}\right), ..., \mathbf{f}\left(\theta_\mathbf{D}\right)\right]$ is the array manifold matrix, and $\mathbf{n}\left[m\right]$ is the white noise vector with power σ_0^2. We assume that the sources are temporally uncorrelated. Hence the autocorrelation matrix of $\boldsymbol{\gamma}\left[m\right]$ is diagonal. Then, the autocorrelation matrix of the received signal is

given by [55]

$$\begin{aligned}\boldsymbol{\Theta}_{\mathbf{rr}} &= \mathbb{E}\left[\mathbf{r}\mathbf{r}^H\right] = \mathbf{F}\boldsymbol{\Theta}_{\boldsymbol{\gamma}\boldsymbol{\gamma}}\mathbf{F}^H + \sigma_0^2\mathbf{I} \\ &= \mathbf{F}\begin{pmatrix} \sigma_1^2 & & & \\ & \sigma_2^2 & & \\ & & \ddots & \\ & & & \sigma_D^2 \end{pmatrix}\mathbf{F}^H + \sigma_0^2\mathbf{I}. \end{aligned} \tag{3.2}$$

We next vectorize $\boldsymbol{\Theta}_{\mathbf{rr}}$ and obtain the following vector [55].

$$\mathbf{z} = \mathrm{vec}\left(\boldsymbol{\Theta}_{\mathbf{rr}}\right) = \mathrm{vec}\left[\sum_{i=1}^{D}\sigma_i^2\left(\mathbf{f}\left(\theta_i\right)\mathbf{f}^H\left(\theta_i\right)\right)\right] + \sigma_0^2\vec{\mathbf{1}} = \left(\mathbf{F}^*\odot\mathbf{F}\right)\mathbf{p} + \sigma_0^2\vec{\mathbf{1}}, \tag{3.3}$$

where $\mathbf{p} = [\sigma_1^2, \sigma_2^2, ..., \sigma_D^2]^T$ is the power vector of the D sources. $\vec{\mathbf{1}} = \left[\mathbf{e}_1^T, \mathbf{e}_2^T, ..., \mathbf{e}_N^T\right]^T$, and $\mathbf{e}_i$ is a column vector with 1 at the ith position and 0 at all other positions. Comparing (3.3) with (3.1), we find that $\mathbf{z}$ can be regarded as a signal received at an array with a manifold matrix given as $\mathbf{F}^* \odot \mathbf{F}$, where $\odot$ denotes the Khatri-Rao (KR) product. The corresponding source signal is $\mathbf{p}$ and the noise vector is given as $\sigma_n^2\vec{\mathbf{1}}_n$. Analyzing the manifold matrix $\mathbf{F}^* \odot \mathbf{F}$, we find that the distinct rows of $\mathbf{F}^* \odot \mathbf{F}$ behave like the manifold of an array with antenna positions given by distinct values in the set $\left\{\vec{\boldsymbol{\varepsilon}}_i - \vec{\boldsymbol{\varepsilon}}_j, 1 \le i, j \le N\right\}$, where $\vec{\boldsymbol{\varepsilon}}_i$ is the position vector of the original array. The new array is the difference co-array of the original array [56].

In a difference co-array with antenna positions given in the set $\left\{\vec{\boldsymbol{\varepsilon}}_i - \vec{\boldsymbol{\varepsilon}}_j\right\}$, for all $i, j = 1, 2, ..., N$, it is easy to see that the number of elements in this set is $N(N-1)+1$. Thus, given the original N-antenna array, the maximum DoFs of a difference co-array is

$$\mathrm{DOF}_{\max} = N(N-1). \tag{3.4}$$

We thus conclude that $O(N^2)$ DoFs can be achieved with N antennas by exploiting the second order statistics of the received signal [55]. This opens tremendous opportunities to detect more sources than the number of physical antenna elements.

3.2.2 Nested Array: An Effective Approach to Increase DoF

Based on the difference co-array framework, the nested array was proposed in [55] as an effective solution to the problem of resolving more sources than antenna elements. Nested array is characterized by *non-uniform antenna array placement* and second order statistic processing of the received signal. According to the analysis in (3.3), the difference co-array of a nested array has $O(N^2)$ antenna elements, and thus a nested array can achieve $O(N^2)$ DoFs. Compared to existing methods on increasing DoFs, the nested array approach is easy to implement with reduced overhead and can be applied to more general scenarios since less assumptions are needed for the system model. In addition, the nested array operates in a passive sensing pattern, which only needs to receive source signals. These favorable features make nested array suitable to applications in cellular networks. The implementation and setup process of a nested array are described in [55].

3.2.3 Interference Nulling with Nested Array

An important application of a nested array is its capability of interference nulling. Let $\mathbf{z} = (\mathbf{F}^* \odot \mathbf{F})\,\mathbf{p} + \sigma_{\mathbf{n}}^{\mathbf{2}}\,\vec{\mathbf{1}}_{\mathbf{n}}$ be the equivalent received signal at the difference co-array of the nested array of an SBS. Suppose we apply a beamforming with weight vector $\mathbf{w}$. Then, the resulting signal is given by

$$r' = \mathbf{w}^H \mathbf{z} = \sum_{i=1}^{D} \mathbf{w}^H \left(\mathbf{f}^* \left(\theta_i\right) \otimes \mathbf{f}\left(\theta_i\right)\right) \sigma_i^2 + \sigma_0^2 \mathbf{w}^H \vec{\mathbf{1}}, \tag{3.5}$$

where $\otimes$ denotes the the Kronecker product. In (3.5), r' can be regarded as a weighted sum of $\sigma_i^2, i = 1, 2, ..., D$, and σ_0^2 with wights given as $\mathbf{w}^H \left(\mathbf{f}^* \left(\theta_i\right) \otimes \mathbf{f}\left(\theta_i\right)\right)$ and $\mathbf{w}^H \vec{\mathbf{1}}$, respectively. Define the new beam pattern as

$$B(\theta_i) = \mathbf{w}^H \left(\mathbf{f}^* \left(\theta_i\right) \otimes \mathbf{f}\left(\theta_i\right)\right). \tag{3.6}$$

Thus, the powers of sources from different directions get spatially filtered by $B(\theta_i), i = 1, 2, ..., D$. It is then possible to adjust these new beam patterns so that the antenna array only receives desired signals, while nulling noise and interference signals.

For an SBS with nested array, suppose the directions of its small cell user equipments (SUE) are $\{\delta_l, l = 1, 2, ..., L\}$, and the directions of interfering SUEs and macrocell user equipments (MUE) are $\{\eta_i, i = 1, 2, ..., I\}$. Then, the beam patterns of different directions are expected to be

$$\begin{cases} B(\delta_l) = 1, \ l = 1, 2, ..., L \\ B(\eta_i) = 0, \ i = 1, 2, ..., I. \end{cases} \tag{3.7}$$

According to the expression of $\mathbf{z}$ in (3.3), the beamforming weight vector should satisfy

$$\begin{pmatrix} (\mathbf{f}^*(\delta_1) \otimes \mathbf{f}(\delta_1))^H \\ \vdots \\ (\mathbf{f}^*(\delta_L) \otimes \mathbf{f}(\delta_L))^H \\ (\mathbf{f}^*(\eta_1) \otimes \mathbf{f}(\eta_1))^H \\ \vdots \\ (\mathbf{f}^*(\eta_I) \otimes \mathbf{f}(\eta_I))^H \\ (\vec{\mathbf{1}})^T \end{pmatrix} \mathbf{w} = \begin{pmatrix} 1 \\ \vdots \\ 1 \\ 0 \\ \vdots \\ 0 \\ 0 \end{pmatrix}. \tag{3.8}$$

With the solution of $\mathbf{w}$, the weight vector of the original antenna array can be determined by the method presented in [55]. It can be observed from (3.8) that the number of DoFs to identify and manage the desired signals, noise, and interfering signals is $O(N^2)$. This enforces a constraint on the number of desired and interference sources that can be managed, and we will consider this in the problem formulation. The nested array based interference nulling approach provides a new perspective to interference management, by employing spatial filtering on different directions. With such desirable features of nested array, it is highly promising to apply this technique to interference management in massive MIMO HetNets.

3.3 System Model and Problem Formulation

We consider a two-tier massive MIMO HetNet consists of one MBS with a massive MIMO (labeled as BS $j = 0$) and multiple SBS's with a regular MIMO (denoted as $j = 1, 2, ..., J$). There are K users (indexed by $k = 1, 2, ..., K$) to be served. Let $x_{k,j}$ be the *user association variable* defined as

$$x_{k,j} \doteq \begin{cases} 1, & \text{user } k \text{ is associated with BS } j \\ 0, & \text{otherwise,} \end{cases} \quad k = 1, 2, ..., K,\ j = 0, 1, ..., J, \tag{3.9}$$

The macrocell and small cells share the same spectrum band and both tiers adopt the time division duplex (TDD) mode in a synchronized way, i.e., the two tiers use the same time period for uplink or downlink transmissions. The SBS's use the nested array to perform interference nulling, so that the uplink interference from a certain number of users can be nulled with the beamforming process presented in (3.8). Since the nested array requires second order processing on all antennas, the MBS with massive MIMO adopts the traditional linear array for DoA estimation and interference management due to complexity concerns. In addition, the DoF of MBS antenna array is sufficient even without nested array due to the large number of antennas. With the DoAs of the interference signals, both the MBS and SBS's can optimize direction of departure (DoD) of their transmissions with analog domain beamforming to avoid downlink interference to a certain number of users. This way, the mutual interference between the BS's and some users can be eliminated.

Define binary variables $n_{k,j}$ as *interference nulling indicators*, given as

$$n_{k,j} \doteq \begin{cases} 1, & \text{BS } j \text{ nulls interference from user } k \\ 0, & \text{otherwise,} \end{cases} \quad k = 1, 2, \ldots, K,\ j = 0, 1, \ldots, J. \tag{3.10}$$

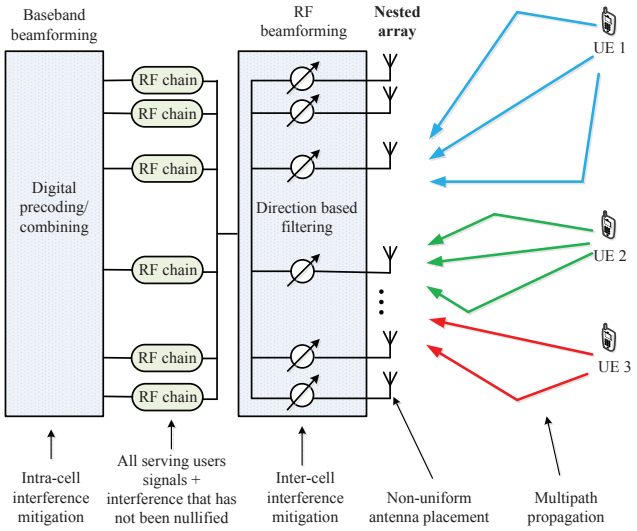

Figure 3.1: System architecture and signal processing of nested array-based interference management.

Due to the multipath propagation, the signal of each user is received by a BS from multiple directions. Thus, the DoF assigned to a user is determined by the number of multipath components between the user and the BS. Based our analysis in (3.8), $n_{k,j}$ should satisfy

$$\sum_{k=1}^{K} x_{k,j} q_{k,j} + \sum_{k=1}^{K} n_{k,j} q_{k,j} + 1 \leq D_j, \ j = 0, 1, ..., J, \tag{3.11}$$

where $q_{k,j}$ is the number of multipath from user k to BS j, D_j is the DoF of BS j, which serves as the upper bound for the number of directions that can be resolved by BS j. Let M_j be the number of antennas at BS j, we assume that the SBS's adopt the optimal N-level nested array, then $D_j = M_j(M_j - 1) + 1$, $j = 1, ..., J$. For the MBS without nested array, we have $D_0 = M_0$. We also assume that noise is always nulled at the BS's with one DoF.

The transmission/reception at each BS is a two-stage process, one is analog processing at the radio frequency (RF) domain and the other is digital processing at baseband, as shown in Fig. 3.1. As presented in Section 3.2, the nested array based interference nulling is performed at the RF domain of each SBS. By identifying the directions of desired signals and interference, the signals from different directions are spatially filtered with the second order processing given in (3.8). Then, the signals of users served by an SBS remain while part of the interference from

other users are nullified. After interference nulling using nested array, the intra-cell interference between users served by the same BS still exists and can only be mitigated with baseband processing. For SBS's, we assume that a matched filter is adopted for precoding and detection. Then, the sum of uplink and downlink data rates of user k connecting to SBS j can be approximated as in (3.12),

$$R_{k,j} = \log\left(1 + \frac{p_k\left|h_{k,j}^H h_{k,j}\right|^2}{\sum\limits_{k'\neq k} x_{k',j}p_{k'}\left|h_{k,j}^H h_{k',j}\right|^2 + \sum\limits_{j'\neq j}\sum\limits_{k'\neq k} x_{k',j'}p_{k'}g_{k',j}\left(1-n_{k',j}\right)}\right)$$
$$+ \log\left(1 + \frac{p_j\left|h_{k,j}^H h_{k,j}\right|^2}{1+\sum\limits_{k'\neq k} x_{k',j}p_j\left|h_{k,j}^H h_{k',j}\right|^2 + \sum\limits_{j'\neq j} p_{j'}g_{k,j'}\left(1-n_{k,j'}\right)}\right),$$
$$k = 1,2,...,K,\ j = 1,...,J. \tag{3.12}$$

where p_k and p_j are the powers of user k and BS j, respectively; $h_{k,j}$ is the channel gain between user k and BS j, and $g_{k,j}$ is the large-scale channel power gain between user k and BS j.

For a macrocell user, due to the law of large numbers, the interference caused by other macrocell users can be averaged out in a massive MIMO system. Using the data rate model of massive MIMO HetNet in [12], the sum of uplink and downlink data rates of user k connecting to MBS is given by (3.13),

$$R_{k,0} = \log\left(1 + \frac{M_0 - S_0 + 1}{S_0}\frac{p_k g_{k,0}}{\sum\limits_{j=1}^{J}\sum\limits_{k'\neq k} x_{k',j}p_{k'}g_{k',0}\left(1-n_{k',0}\right)}\right)$$
$$+ \log\left(1 + \frac{M_0 - S_0 + 1}{S_0}\frac{p_0 g_{k,0}}{1+\sum\limits_{j=1}^{J} p_j g_{k,j}\left(1-n_{k,j}\right)}\right),$$
$$k = 1,2,...,K. \tag{3.13}$$

where M_0 is the number of antennas of MBS, S_0 is the beamforming size of MBS, $\frac{M_0-S_0+1}{S_0}$ is the antenna array gain of massive MIMO.

Let $\mathbf{x}$ and $\mathbf{n}$ be the matrices of $\{x_{k,j}\}$ and $\{n_{k,j}\}$, respectively. The sum rate maximization of a massive MIMO HetNet is formulated as follows.

$$\mathbf{P1}: \max_{\{\mathbf{x},\mathbf{n}\}} \left\{ \sum_{k=1}^{K} x_{k,0} R_{k,0} + \sum_{k=1}^{K} \sum_{j=1}^{J} x_{k,j} R_{k,j} \right\} \tag{3.14}$$

subject to:

$$\sum_{k=1}^{K} x_{k,j} q_{k,j} + \sum_{k=1}^{K} n_{k,j} q_{k,j} + 1 \leq D_j, \ j = 0, 1, ..., J \tag{3.15}$$

$$n_{k,j} \leq 1 - x_{k,j}, \ k = 1, 2, \ldots, K, \ j = 0, 1, ..., J \tag{3.16}$$

$$\sum_{j=0}^{J} x_{k,j} \leq 1, \ k = 1, 2, \ldots, K \tag{3.17}$$

$$x_{k,j}, n_{k,j} \in \{0, 1\}, \ k = 1, 2, \ldots, K, \ j = 0, 1, ..., J. \tag{3.18}$$

Constraint (3.16) is due to the fact that when BS j serves user k, it does not need to null interference from user k. Constraint (3.17) is the constraint on user association.

3.4 Interference Management with a Given User Association

In this section, we consider the case that the set of users served by each BS is pre-determined (e.g., through a user association algorithm [13]). Then, problem **P1** is transformed to the following problem.

$$\mathbf{P2}: \max_{\{\mathbf{n}\}} \left\{ \sum_{k=1}^{K} x_{k,0} R_{k,0} + \sum_{k=1}^{K} \sum_{j=1}^{J} x_{k,j} R_{k,j} \right\} \tag{3.19}$$

subject to:

$$\sum_{k=1}^{K} x_{k,j} q_{k,j} + \sum_{k=1}^{K} n_{k,j} q_{k,j} + 1 \leq D_j, \ j = 0, 1, ..., J \tag{3.20}$$

$$n_{k,j} \leq 1 - x_{k,j}, \ k = 1, 2, \ldots, K, \ j = 0, 1, ..., J \tag{3.21}$$

$$n_{k,j} \in \{0, 1\}, \ k = 1, 2, \ldots, K, \ j = 0, 1, ..., J. \tag{3.22}$$

Problem **P2** is an integer programming program with a nonlinear and non-convex objective function, which is generally NP-hard. To make the problem tractable, we assume the system operate in the high SINR regime, so that $\log(1+\mathrm{SINR}) \approx \log(\mathrm{SINR})$. The high SINR assumption is reasonable in a massive MIMO HetNet due to the large antenna array gain of massive MIMO and the short transmission distance of small cells. Applying this approximation to (3.12) and (3.13), the objective function of problem **P2** can be written as

$$\sum_{k=1}^{K}\sum_{j=0}^{J} x_{k,j}V_{k,j},\ k=1,2,...,K,\ j=0,1,...,J, \tag{3.23}$$

$V_{k,j}$ is given as

$$\begin{aligned} V_{k,0} = &\log\left(\frac{M_0-S_0+1}{S_0}p_k g_{k,0}\right) + \log\left(\frac{M_0-S_0+1}{S_0}p_0 g_{k,0}\right) \\ &- \log\left(\sum_{j=1}^{J}\sum_{k'\neq k} x_{k',j}p_{k'}g_{k',0}\left(1-n_{k',0}\right)\right) - \log\left(1+\sum_{j=1}^{J} p_j g_{k,j}\left(1-n_{k,j}\right)\right) \end{aligned} \tag{3.24}$$

$$\begin{aligned} V_{k,j} = &\log\left(p_k g_{k,j}\right) + \log\left(p_j g_{k,j}\right) - \log\left(I_{k,j}^{U} + \sum_{j'\neq j}\sum_{k'\neq k} x_{k',j'}p_{k'}g_{k',j}\left(1-n_{k',j}\right)\right) \\ &- \log\left(1+I_{k,j}^{D} + \sum_{j'\neq j} p_{j'}g_{k,j'}\left(1-n_{k,j'}\right)\right), \end{aligned} \tag{3.25}$$

where $I_{k,j}^{U} = \sum\limits_{k'\neq k} x_{k',j}p_{k'}\left|h_{k,j}^{H}h_{k',j}\right|^2$ and $I_{k,j}^{D} = \sum\limits_{k'\neq k} x_{k',j}p_{j}\left|h_{k,j}^{H}h_{k',j}\right|^2$.

Let $\mathcal{U}_j$ be the set of users served by BS j, $\mathcal{U}_j = \{k \,|x_{k,j}=1\}$. Then, the objective function of **P2** can be rewritten as $\sum_{k\in\mathcal{U}_j}\sum_{j=0}^{J} x_{k,j}V_{k,j}$. We remove the constants in (3.24) and (3.25) and apply the property $\sum_i \log x_i = \log\left(\prod_i x_i\right)$. Since $\log(\cdot)$ is a monotonic function, **P2** can be transformed into the following problem.

$$\mathbf{P3}:\ \max_{\{\mathbf{n}\}} \prod_{j=0}^{J}\prod_{k\in\mathcal{U}_j} W_{k,j} \tag{3.26}$$

$$\text{subject to: } (3.20), (3.21), \text{and } (3.22),$$

where

$$W_{k,0} = \left[\sum_{j=1}^{J}\sum_{k'\neq k} x_{k',j}p_{k'}g_{k',0}\left(1-n_{k',0}\right)\right] \times \left[1+\sum_{j=1}^{J} p_j g_{k,j}\left(1-n_{k,j}\right)\right], \tag{3.27}$$

$$W_{k,j} = \left[I_{k,j}^{U} + \sum_{j'\neq j}\sum_{k'\neq k} x_{k',j'}p_{k'}g_{k',j}\left(1-n_{k',j}\right)\right] \times \left[I_{k,j}^{D} + 1 + \sum_{j'\neq j} p_{j'}g_{k,j'}\left(1-n_{k,j'}\right)\right] \tag{3.28}$$

It can be seen that the objective function of **P3** is a product of linear expressions, which can be expressed as a polynomial on the set of variables $\{n_{k,j}\}$. Thus, **P3** is a nonlinear integer programming problem with a complicated form, which is hard to solve with normal approaches. However, we can make use of a property of 0-1 problems to approximate problem **P3** with a linear integer programming problem. Then, the cutting plane method [57] can be employed to effectively obtain the optimal solution to the linear integer programming problem.

3.4.1 Linear Approximation of **P3**

Consider the product of multiple i.i.d. 0-1 variables. When the number of variables is increased, the product becomes less likely to be 1 since it is less likely that all the variables are 1. As the objective function of **P2** is a weighted sum of products of 0-1 variables, the values of higher-order parts are more likely to be 0. Thus, the impact of the higher-order parts is limited. Let P be the probability that an arbitrary $n_{k,j}$ equals to 1. In the objective function of **P3**, the probability for an M-th order product to be 1 is P^M.

Lemma 13 *P can be approximated by $\frac{\bar{D}_j-1}{\bar{q}_{k,j}K} - \frac{1}{J} - \frac{1}{\bar{q}_{k,j}JK}$, where $\bar{z}$ is the mean of a variable z.*

Proof: To maximize the sum rate, all DoFs of each BS are expected to be used for data transmission and interference nulling. Thus, all the constraints described by (3.20) are close to equality. Adding these equations from $j = 0$ to $j = J$, we have $\sum_{k=1}^{K}\sum_{j=0}^{J} n_{k,j}q_{k,j} =$

$\sum_{j=0}^{J} D_j - \sum_{k=1}^{K}\sum_{j=0}^{J} x_{k,j}q_{k,j} - J - 1$. The probability that $n_{k,j}$ equals to 1 can be derived as

$$P = \frac{\sum\limits_{k=1}^{K}\sum\limits_{j=0}^{J} n_{k,j}q_{k,j}}{\sum\limits_{k=1}^{K}\sum\limits_{j=0}^{J} q_{k,j}} = \frac{\sum\limits_{j=0}^{J} D_j - \sum\limits_{k=1}^{K}\sum\limits_{j=0}^{J} x_{k,j}q_{k,j} - J - 1}{\bar{q}_{k,j}JK}$$
$$= \frac{\sum\limits_{j=0}^{J} D_j - \bar{q}_{k,j}K - J - 1}{\bar{q}_{k,j}JK} = \frac{\bar{D}_j - 1}{\bar{q}_{k,j}K} - \frac{1}{J} - \frac{1}{\bar{q}_{k,j}JK}.$$

■

In a typical cellular network, the number of users in a macrocell can be more than 500, i.e., $K > 500$. The DoFs are $D_j = O(N^2)$ for SBS's and $D_0 = O(N)$ for the MBS. As the typical number of antennas for MBS and SBS are 100 and 10, respectively, we have $\bar{D}_j \approx 100$. Therefore, the value of P is expected to be small in a practical system, and the values of higher-order terms of P are close to 0. Due to this fact, we derive linear approximations for the higher-order parts in the polynomial of (3.26) and transform the objective of **P3** to a linear function.

Let $\tilde{\mathbf{n}}$ be the vector concatenating the columns of matrix $[n_{k,j}]_{K\times J}$. For a product with M elements in $\tilde{\mathbf{n}}$ given as $\tilde{\mathbf{n}}_{i_1}, \tilde{\mathbf{n}}_{i_2}, \cdots \tilde{\mathbf{n}}_{i_M}$, we have the following approximation.

$$\tilde{\mathbf{n}}_{i_1}\tilde{\mathbf{n}}_{i_2}\cdots\tilde{\mathbf{n}}_{i_M} \approx \frac{P^{M-1}}{M}\left(\tilde{\mathbf{n}}_{i_1} + \tilde{\mathbf{n}}_{i_2} + \cdots + \tilde{\mathbf{n}}_{i_M}\right). \tag{3.29}$$

It can be easily verified that the expectations of both sides are equal to P^M, thus the long term performance of the approximation problem equals to that of the original problem. The expected value of the gap between the two sides of (3.29) is given as

$$\mathbb{E}\left\{\tilde{\mathbf{n}}_{i_1}\tilde{\mathbf{n}}_{i_2}\cdots\tilde{\mathbf{n}}_{i_M} - \frac{P^{M-1}}{M}\left(\tilde{\mathbf{n}}_{i_1} + \tilde{\mathbf{n}}_{i_2} + \cdots + \tilde{\mathbf{n}}_{i_M}\right)\right\} = (M-1)P^M,$$

which is a quite small value even when M is small. In addition, the product on the left hand side of (3.29) is approaching 0 as M increases. Thus, the approximation given by (3.29) is expected to be accurate.

With the linear approximation of the polynomial objective function, **P3** is transformed to the following integer programming problem.

$$\mathbf{P4}: \max_{\{\tilde{\mathbf{n}}\}} \ \mathbf{c}\tilde{\mathbf{n}} \tag{3.30}$$

$$\text{subject to: } \ \mathbf{A}\tilde{\mathbf{n}} \leq \mathbf{b}. \tag{3.31}$$

The vector $\mathbf{c}$ is determined by applying the linear transformation of (3.29) to (3.26). The constraint matrix $\mathbf{A}$ is given by

$$\mathbf{A}_{(K+1)(J+1)\times K(J+1)} \doteq \begin{pmatrix} \mathbf{Q} \\ \mathbf{I} \end{pmatrix}, \tag{3.32}$$

where $\mathbf{I}$ is a $K\left(J+1\right) \times K\left(J+1\right)$ identity matrix, and $\mathbf{Q}$ is given by

$$\mathbf{Q}_{(J+1)\times K(J+1)} = \begin{pmatrix} \mathbf{q}_0 & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{q}_1 & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{q}_J \end{pmatrix}, \tag{3.33}$$

where $\mathbf{q}_j = \left[q_{1,j}, q_{2,j}, ..., q_{K,j}\right], \ j = 0, 1, ..., J$. The vector $\mathbf{b}$ in (3.31) is given by

$$\mathbf{b}_{(K+1)(J+1)\times 1} \doteq \left[E_0, ..., E_J, 1 - x_{1,0}, ..., 1 - x_{K,0}, 1 - x_{1,1}, ..., 1 - x_{K,J}\right]^T, \tag{3.34}$$

where

$$E_j = D_j - \sum_{k=1}^{K} x_{k,j} q_{k,j} - 1, \ j = 0, 1, ..., J. \tag{3.35}$$

In **P4**, matrix $\mathbf{A}$ characterizes the coefficients of the linear constraints (3.20) and (3.21) on $\{n_{k,j}\}$. In $\mathbf{A}$, the matrix $\mathbf{Q}$ corresponds to the information on the number of multipath between each user and each BS. The vector $\mathbf{b}$ consists of the values of the right-hand side of constraints

(3.20) and (3.21). In particular, $E_0, ..., E_J$ are the upper bounds for the DoF that can be used by the BS's.

3.4.2 Performance Upper Bound

To verify the effectiveness of the approximation, we derive a performance upper bound for **P3** and compare it with the proposed scheme in our simulations. Apply the following linear approximation

$$\tilde{\mathbf{n}}_{i_1}\tilde{\mathbf{n}}_{i_2}\cdots\tilde{\mathbf{n}}_{i_M} \approx \frac{\tilde{\mathbf{n}}_{\mathbf{i_1}} + \tilde{\mathbf{n}}_{i_2} + \cdots + \tilde{\mathbf{n}}_{i_M}}{M}. \tag{3.36}$$

The resulting integer programming can be expressed as

$$\mathbf{P5}: \max_{\tilde{\mathbf{n}}} \ \mathbf{c}'\tilde{\mathbf{n}} \tag{3.37}$$

$$\text{subject to: } \mathbf{A}\tilde{\mathbf{n}} \leq \mathbf{b}, \tag{3.38}$$

where $\mathbf{c}'$ is determined by (3.36).

Lemma 14 *With the linear approximation described in* (3.36)*, the objective function of problem* **P5** *is an upper bound for that of problem* **P3***.*

Proof: Due to the fact that the geometric mean is no greater than the arithmetic mean, we have $\sqrt[M]{\tilde{\mathbf{n}}_{\mathbf{i_1}}\tilde{\mathbf{n}}_{i_2}\cdots\tilde{\mathbf{n}}_{i_M}} \leq (\tilde{\mathbf{n}}_{i_1} + \tilde{\mathbf{n}}_{i_2} + \cdots + \tilde{\mathbf{n}}_{i_M})/M$. Since all elements of $\tilde{\mathbf{n}}$ are 0-1 variables, it can be easily verified that $\sqrt[M]{\tilde{\mathbf{n}}_{i_1}\tilde{\mathbf{n}}_{i_2}\cdots\tilde{\mathbf{n}}_{i_M}} = \tilde{\mathbf{n}}_{i_1}\tilde{\mathbf{n}}_{i_2}\cdots\tilde{\mathbf{n}}_{i_M}$. Thus, we have

$$\tilde{\mathbf{n}}_{i_1}\tilde{\mathbf{n}}_{i_2}\cdots\tilde{\mathbf{n}}_{i_M} \leq \frac{\tilde{\mathbf{n}}_{i_1} + \tilde{\mathbf{n}}_{i_2} + \cdots + \tilde{\mathbf{n}}_{i_M}}{M}, \ \forall M \geq 2. \tag{3.39}$$

Applying (3.39) to all the higher-order expressions in (3.26), $\mathbf{c}'\tilde{\mathbf{n}}$ is an upper bound to the objective function of **P3** ■

Based on Lemma 14, we further conclude that the optimal solution to problem **P5** provides an upper bound to the optimal solution of **P3**.

3.4.3 Optimal Solution to **P4** with the Cutting Plane Method

Since problem **P4** has a linear objective function, the cutting plane method [57] can be used to derive the optimal solution. The idea of cutting plane is to find a plane that separates the non-integer solution from the polyhedron that satisfies the constraints and contains all the integer, feasible solutions.

Consider the polyhedron defined by $\mathbf{A}\tilde{\mathbf{n}} \leq \mathbf{b}$, determined by a combination of $(K+1)(J+1)$ inequalities as

$$\mathbf{a}_i\tilde{\mathbf{n}} \leq b_i,\ i = 1, 2, ..., (K+1)(J+1), \tag{3.40}$$

where $\mathbf{a}_i$ is the ith row of $\mathbf{A}$. Let $y_1, y_2, ..., y_{(K+1)(J+1)} \geq 0$ and set

$$\mathbf{a}^* = \sum_{i=1}^{(K+1)(J+1)} y_i\mathbf{a}_i,\quad b^* = \sum_{i=1}^{(K+1)(J+1)} y_i b_i. \tag{3.41}$$

Obviously, all solutions in the polyhedron $\mathbf{A}\tilde{\mathbf{n}} \leq \mathbf{b}$ also satisfy $\mathbf{a}^*\tilde{\mathbf{n}} \leq b^*$. If $\mathbf{a}^*$ is integral, i.e., all elements of $\mathbf{a}^*$ are integers, then all the *integer* solutions should satisfy

$$\mathbf{a}^*\tilde{\mathbf{n}} \leq \lfloor b^* \rfloor, \tag{3.42}$$

where $\lfloor b^* \rfloor$ is the largest integer that is smaller than b^*. Then, (3.42) defines a cutting plane for **P4**.

With an additional constraint described by the cutting plane, all the integer solutions are still included while some non-integer solutions are removed. Due to this property, we can first relax the integer constraint in **P4** and solve the linear programming problem. If there are non-integer solutions, we add an additional constraint in the form of (3.42). Then, we solve the linear programming problem with the updated constraints. If there are still non-integer elements in the solution vector, we continue to add another constraint following (3.42). Such process terminates when all solution variables are integer. However, the effectiveness of this approach

depends on the proper setting of parameters y_i. An efficient scheme to find the effective cutting plane was proposed in [57]; the details are omitted here due to lack of space.

3.4.4 A Special Case without the Need for Cutting Plane

We consider a special case with an additional condition.

Assumption 1 *Each BS j only receives a fixed amount of L_j strongest multipath signals from each user and neglect the other multipath components with weaker signal strengths, and L_j is set to a be value such that E_j/L_j is an integer, where E_j is defined in* (3.35)*.*

With Assumption 1, we then have

$$q_{1,j} = q_{2,j} = \cdots = q_{K,j} = L_j,\ j = 0, 1, ..., J. \tag{3.43}$$

Note that, when L_j is sufficiently large, this special case can be regarded as the real case. Given (3.41), we can divide both sides of (3.20) by L_j. Then, $\mathbf{A}$ is updated by replacing $\mathbf{q}_j$ with

$$\mathbf{q}_j = [1, 1, ..., 1]^T. \tag{3.44}$$

The vector $\mathbf{b}$ is updated as

$$\mathbf{b}_{(K+1)(J+1)\times 1} \doteq \left[\frac{E_0}{L_0}, ..., \frac{E_J}{L_J}, 1 - x_{1,0}, ..., 1 - x_{K,0}, 1 - x_{1,1}, ..., 1 - x_{K,J}\right]^T. \tag{3.45}$$

It can be seen that there are exactly two 1's in each column of $\mathbf{A}$, with one from a column of $\mathbf{Q}$ and the other from a column of $\mathbf{I}$.

Definition 1 *A matrix $\mathbf{A}$ is totally unimodular if the determinant of every square submatrix of $\mathbf{A}$ is either 0, +1 or -1 [44].*

Lemma 15 *Under Assumption 1, $\mathbf{A}$ is a totally unimodular matrix.*

Proof: We divide the constraint matrix $\mathbf{A}$ into blocks as

$$\mathbf{A} = \begin{pmatrix} \mathbf{Q}_1 & \mathbf{Q}_2 & \cdots & \mathbf{Q}_J \\ \mathbf{I}_1 & \mathbf{I}_2 & \cdots & \mathbf{I}_J \end{pmatrix},$$

where each $\mathbf{Q}_j$, $j = 1, 2, \ldots, J$, is a $J \times K$ matrix; the j th row of $\mathbf{Q}_j$ is all 1, while all the other rows are all 0; and each $\mathbf{I}_j$, $j = 1, 2, \ldots, J$, is a $K \times K$ identity matrix.

Denote G_n as an arbitrary $n \times n$ square submatrix of matrix $\mathbf{A}$. Obviously, the determinant of G_n is either 0 or 1 when $n = 1$. To analyze the determinant of G_n for $n \geq 2$, the following two cases need to be considered.

Case 1: G_n is a submatrix of $\mathbf{Q}_j$ or $\mathbf{I}_j$, $j = 1, 2, \ldots, J$. If G_n is a submatrix of $\mathbf{Q}_j$, we have $\det(G_n) = 0$, since at least one row would be all 0. If G_n is a submatrix of $\mathbf{I}_j$, $\det(G_n)$ would be either 0 or $+1$, since $\mathbf{I}_j$ is an identity matrix.

Case 2: The entries of G_n are from more than one $\mathbf{Q}_j$ or $\mathbf{I}_j$. We apply an induction method to analyze the determinant. For $n = 2$, $\det(G_n)$ can only be 0, $+1$, or -1, since the four entries are either 0 or 1 with at least one 0. Suppose $\det(G_{n-1})$ can only be 0, $+1$, or -1, we need to verify whether the same result hold for $\det(G_n)$. Denote $G_n(u, v)$ as the entry of G_n at row u, column v. Let $v^* = \arg\min_v \{\sum_u G_n(u, v)\}$. Then column v^* is the one with the minimum number of 1s in G_n. Let φ_{v^*} be the number of 1s in column v^*, which can be 0, 1, or 2 according to the structure of $\mathbf{A}$.

If $\varphi_{v^*} = 0$, column v^* of G_n is all 0 and $\det(G_n) = 0$.

If $\varphi_{v^*} = 1$, we calculate $\det(G_n)$ through column v^* and have $\det(G_n) = \pm \det(G_{n-1})$. According to the induction hypothesis, $\det(G_{n-1})$ can only be 0, -1, or 1. Therefore, $\det(G_n)$ can only be 0, -1, or 1.

If $\varphi_{v^*} = 2$, each column of G_n has exactly two 1s, with one in $\mathbf{Q}_j$ and the other in $\mathbf{I}_j$. Due to the equal number of 1s in $\mathbf{Q}_j$ and $\mathbf{I}_j$, we can obtain an all-zero row in G_n through some elementary transformations, which yields $\det(G_n) = 0$.

Consequently, the determinant of any square submatrix of $\mathbf{A}$ can only be either 0, -1, or 1. According to Definition 1, we conclude that $\mathbf{A}$ is totally unimodular. ■

For a linear programming problem with a unimodular constraint matrix $\mathbf{A}$ and integral right hand side vector $\mathbf{b}$, all decision variables to the optimal solution are integers [44]. Thus, the optimal solution of **P4** can be obtained by relaxing the integer constraints and solving the resulting linear programming problem.

3.5 Distributed Algorithm for Joint Interference Nulling Schedule and User Association

As the DoF of each BS is shared by interference nulling and data transmission of users, joint optimization of interference nulling schedule and user association, which corresponds to solving problem **P1**, could further optimize the system performance. However, problem **P1** is a highly complicated integer programming problem with two sets of variables, which cannot be solved with computational efficient techniques. In this section, we propose a distributed solution algorithm based on a poly matching between users and BS's, in which each user and BS makes its own decision to optimize its performance.

3.5.1 Poly Matching Between Users and BS's

We assume that each user has a *preference list* over the BS's. When a user is not served by any BS, the preference list is for user association, which is determined by the achievable rate of connecting to different BS's. For instance, if $j^* = \arg\max_j \{R_{k,j}\}$, BS j^* is on top of user k's preference list. When, a user is served by a BS, the preference list is for interference nulling, which is determined by the level of interference received from other BS's.

On the other hand, each BS has a preference list over users, which is determined by its performance gain achieved by serving a user or nullifying the interference of a user. Each BS also has a waiting list indicating the set of users that are currently hold by the BS. With the objective of maximizing the sum rate of the users that it serves under the constraint $\sum_{k=1}^{K} x_{k,j} q_{k,j} + \sum_{k=1}^{K} n_{k,j} q_{k,j} + 1 \leq D_j$, the distributed user association and interference nulling strategy for BS j is presented in Algorithm 3.

In Algorithm 3, we define $\Delta_{k,j}$ as the performance gain of BS j by serving user k or nullifying the interference from user k. Suppose BS j put user k^* into its waiting list, then

Algorithm 3: Distributed User Association and Interference Nulling Strategy of BS j

while (*covergence not achieved*) **do**
  For the users propose to BS j and the users that are not served but can be detected by BS j, put them into multiple sets according to the number of multipaths to BS j, given as $\Omega_q = \{k \,|q_{k,j} = q\}$;
  for $q = q_{\min} : q_{\max}$ **do**
    Assign indices to users in Ω_q, $k^* = 1, 2, ...$, according to the descending order of $\Delta_{k,j}$;
    while (*user k^* has not been rejected*) **do**
      if $\sum_{k=1}^{K} x_{k,j} q_{k,j} + \sum_{k=1}^{K} n_{k,j} q_{k,j} + 1 > D_j$ *after accepting user* k^* **then**
        For users already in the waiting list, sort $\{\Delta_{k,j}\}$ is ascending order as $\{\Delta_{i,j}\}$;
        $\rho_j = 0$;
        $i = 1$;
        while $\rho_j < q$ **do**
          $\rho_j = \rho_j + \Delta_{i,j}$;
          $i = i + 1$;
        end
        if $\Delta_{k^*,j} > \Delta_{\{1,...,i\},j}$ **then**
          Add user k^* into the waiting list ;
          Reject user(s) $\{1, ..., i\}$;
          $k^* = k^* + 1$;
        else
          Reject user k^* ;
        end
      else
        Add user k^* in the waiting list ;
        $k^* = k^* + 1$;
      end
    end
    Reject all users ranked after user k^* in Ω_q ;
  end
end

$\Delta_{k^*,j}$ is given as

$$\Delta_{k^*,j} = \sum_{k=1}^{K} x^*_{k,j} R^*_{k,j} - \sum_{k=1}^{K} x_{k,j} R_{k,j}, k = 1, 2, ..., K, \; j = 0, 1, ..., J, \tag{3.46}$$

where $R^*_{k,j}$ and $x^*_{k,j}$ are data rate and user association indicator of user k after serving or nulling the interference of user k^*, respectively. Similarly, $\Delta_{\{1,...,i\},j}$ is defined as the performance gain of BS j by adding the set of users $\{1, ..., i\}$ into its waiting list. The initial values of $\Delta_{k,j}$ are set to be $R_{k,j}$, and $\Delta_{k,j}$ is updated whenever user k proposes to a BS or BS j accepts the proposal of a user. In Algorithm 3, the information needed to calculate $\Delta_{k,j}$, e.g., channel gain and traffic

load, is collected by each BS using the uplink signals from users, and each BS distributively updates its decision variables based on $\Delta_{k,j}$.

The poly matching between users and BS's has three stages.

In the *first* stage, each user proposes to the top BS in its preference list. In particular, if a user has not been served by any BS, the user proposes to be served; for users that are currently served by a BS, they propose to other BS's for interference nulling.

In the *second* stage, the BS's decide whether to accept the proposals of users according to Algorithm 3 and feedback the decision to users. Specifically, the evaluation of each BS begins from the users with the least number of multipaths, to the users with larger numbers of multipaths. For users in each Ω_q, where $\Omega_q = \{k \,|q_{k,j} = q\}$, the evaluation is performed in descending order of $\Delta_{k,j}$. When evaluating a user, if the DoF constraint is still satisfied after serving the user or nullifying the interference of the user, the user is directly put into the waiting list of the BS. If the DoF constraint is violated after adding the user into the waiting list, a BS first selects the user(s) that use a DoF of no less than the required DoFs of the requesting user and with the least performance gain. Then, the BS compares the performance gain of the selected user(s) with the performance gain of the new user, and the one with a larger performance gain will be added or kept in the waiting list.

In the *third* stage, a user that has been rejected first deletes the BS that rejected it from its preference list. Then, the user proposes to the most desirable BS among the remaining ones. After receiving a proposal, a BS compares it with users in its waiting list by evaluating the number of multipath and the achievable rate of the user according to Algorithm 3, and then makes decisions on whether to serve the user. If a user is rejected again, it continues to propose to other BS's following the preference order, and the BS's make decisions and feedback to users, and so forth. Such a matching process between users and BS's regarding user association is continued until convergence, i.e., the users in the waiting list of each BS do not change anymore.

The complexity of the poly matching is upper bounded by $J \cdot K$, which corresponds to the case that every user has proposed to every BS. Since $\Delta_{k,j}$ is updated whenever user k proposes to a BS or BS j accepts the proposal of a user, the complexity of calculating $\Delta_{k,j}$ is also upper

bounded by $J \cdot K$. In Algorithm 3, users with the same number of multipaths are put into set Ω_q. In each set, users are sorted following the descending order of $\Delta_{k,j}$, and then users are evaluated in such order until a user is rejected. When a user in a set is rejected after evaluation by a BS, all subsequent users in the same set will also be rejected since their performance gains are smaller while they have the same number of multipaths. Thus, the BS does not need to evaluate the remaining users and it can directly switch to users in another set. This way the complexity is significantly reduced.

3.5.2 Convergence Analysis

We next prove that the poly matching scheme converges and a stable matching can be achieved.

Definition 2 *In a stable matching, there is no such pair of people who are not matched as partners, while both of them prefer each other to their current partners. In other words, there is no such a pair of people that both of them have a better choice than their current partners [126].*

Lemma 16 *By Algorithm 3, a user entering the waiting list of a BS has a larger performance gain compared to user(s) in the waiting list. In other words, the sum rate of each BS is always increased after adding a user to its waiting list.*

Proof: In Algorithm 3, if the inequality $\sum_{k=1}^{K} x_{k,j} q_{k,j} + \sum_{k=1}^{K} n_{k,j} q_{k,j} + 1 > D_j$ holds after accepting a new user with q multipaths, a comparison would be performed between the incoming user with the user(s) already in the list. The one with larger performance gain would win the competition. Thus, a newly accepted user has a larger performance gain compared to user(s) in the waiting list. We further conclude that the sum rate of users served by a BS is always increased after users are added to the waiting list of the BS. ■

Lemma 17 *The BS's proposed by a user is non-increasing in the user's preference list.*

Proof: Suppose user k is rejected by BS j. According to Algorithm 3, there must be a user k' with $\Delta_{k',j} > \Delta_{k,j}$ and $q_{k',j} \leq q_{k,j}$, and user k' has the smallest value of performance gain among users with a number of multipaths no greater than $q_{k,j}$.

For the case when $q_{k',j} < q_{k,j}$, there is no user in the waiting list of BS j who has a larger or equal number of multipath and a smaller performance gain compared to user k. Hence, BS j would reject user k again, i.e., user k can never enter the waiting list of BS j again. For the case when $q_{k',j} = q_{k,j}$, BS j would make a comparison between user k' and user k. There two subcases: (i) user k' is served by BS j, (ii) BS j nulls the interference of user k'. If user k' is served by BS j, then $\Delta_{k',j} > \Delta_{k,j}$ would always hold regardless of the interference pattern since the SINRs with user k' are always higher than the ones with user k. If user k' is selected by BS j for interference nulling, it is obvious that BS j would reject user k as long as user k' is still in the waiting list. Note that, user k' may be replaced by another user k'' in future rounds. With Lemma 16, we have $\Delta_{k'',j} > \Delta_{k',j} > \Delta_{k,j}$. If user k'' is selected for service provision, we have $\Delta_{k'',j} > \Delta_{k,j}$, user k'' would always bring a higher performance gain than user k regardless of the interference pattern. Thus, user k cannot be accepted by BS j again. If user k'' is selected for interference nulling, it must be the case that user k'' causes stronger interference to other users compared to user k'. Similarly, BS j would reject user k as long as user k'' is still in the waiting list. Thus, it is impossible for user k to be accepted by BS j again. We conclude that the sequence of BS's proposed by a user is non-increasing in its preference list. ■

With Lemma 17, a user deletes a BS from its preference list once it is rejected by the BS.

Theorem 5 *The poly matching converges to a stable matching.*

Proof: We provide a proof by contradiction. Suppose the matching is not stable. By definition, there must be a user k and a BS j such that: (i) user k prefers BS j to its current connecting BS j', (ii) BS j prefers user k to user k', who is currently in the waiting list of BS j and $q_{k',j} \geq q_{k,j}$. Note that, we use the case of one user, i.e., user k' as an example. The proof can be easily extend to the case of multiple users.

Since user k prefers BS j to BS j', user k will propose to BS j. With Algorithm 3, BS j would accept the proposal of user k and replace user k' since it prefers user k over user k'. Based on Lemma 16, the users being put into the waiting list has incremental performance gains. As user k contributes larger performance gain than user k' while user k is not in the

waiting list, it must be the case that user k has never proposed to BS j. As user k prefers BS j over BS j', it must propose to BS j before BS j'. Then, the only explanation is that user k has never proposed to BS j'. However, user k is currently in the waiting list of BS j'. Hence, user k must has proposed to BS j' before, which is a contradiction. We conclude that the poly matching process converges and the outcome is a stable matching. ■

3.6 Simulation Study

We validate the performance of the proposed scheme with Matlab simulations. We consider a macrocell overlaid with multiple small cells. The radii of the macrocell and a small cell are 1000 m and 50 m, respectively. The macrocell and small cells share a total bandwidth of 4 MHz. The transmit power of the MBS is set to 40 dBm, while the transmit power of the MUEs has five levels ranging from 10 dBm to 30 dBm according to the distance between the MUE and MBS. The transmit power of an SBS and an SUE are set to 25 dBm and 15 dBm, respectively. We employ the ITU path loss model [59], the path loss from the MBS to a user and from an SBS to a user are $15.3 + 37.6\log_{10}d$ and $37 + 30\log_{10}d$, respectively. The ratio $\frac{M_0 - S_0 + 1}{S_0}$ is set to 100 for the MBS.

We consider four schemes in our simulations. The first one is interference nulling with a given user association (termed *IN Only*) described in Section 3.4, in which we assume that each user is connected to the BS with the strongest received signal strength. We also evaluate the performance of the proposed distributed joint interference nulling and user association scheme (termed *DJINUA*) presented in Section 3.5. We also consider a heuristic scheme for comparison purpose (termed *Heuristic*). In the heuristic scheme, each user is served by the BS with the strongest signal strength, then each BS chooses the user that causes the strongest interference and nullifies the signals of this user. This process is continued until the DoF constraint at a BS is violated. We also consider the case where no interference nulling is performed as a baseline (termed *No Nulling*).

Two kinds of user distribution patterns are considered, the uniform distribution and the non-uniform distribution. Suppose a total number of U users. In the uniform case, users are

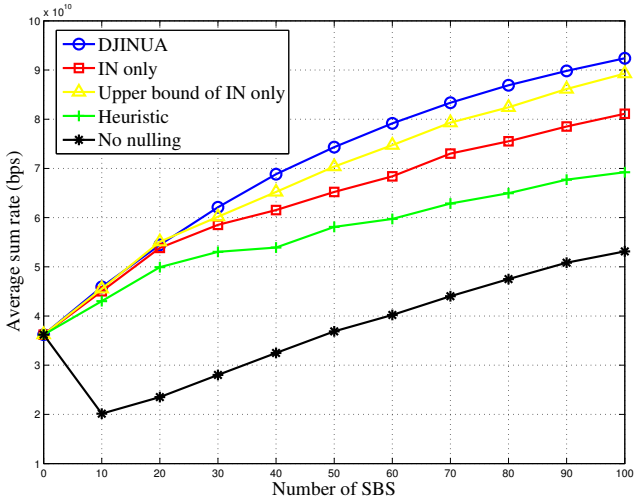

Figure 3.2: Average sum rate versus number of SBS's. Uniform user distribution, 500 users, $\overline{q} = 6$, 10 antennas at each SBS.

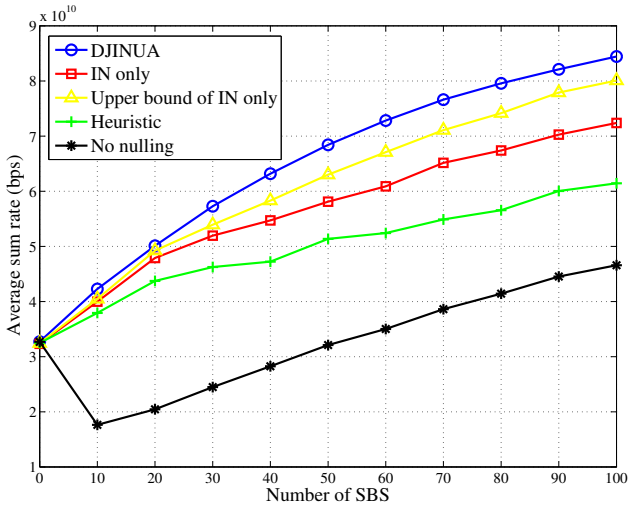

Figure 3.3: Average sum rate versus number of SBS's. Non-uniform user distribution, 500 users, $\overline{q} = 6$, 10 antennas at each SBS.

randomly distributed in the macrocell area; In the non-uniform case, the numbers of users within the coverage of SBS's are random numbers, which generated with the following approach. We first generate J random integers in $[0, U]$. Then, we sort the J integers in ascending order, given as $\kappa_1 \leq \kappa_2 \leq, \cdots, \leq \kappa_J$. Let $\pi_j = \kappa_j - \kappa_{j-1}$, for $j = 2, 3, \cdots, J$, and $\pi_1 = \kappa_1, \pi_{J+1} = U - \kappa_J$. Then, the sequence $\{\pi_1, \pi_2, \cdots, \pi_J\}$ includes $J+1$ random integers in $[0, U]$ and the sum of these integers is U. The number of users in the coverage area of SBS j is set to $\pi_j, j = 1, 2, \cdots, J$, while the number of users that is not in the coverage area of any SBS is π_{J+1}.

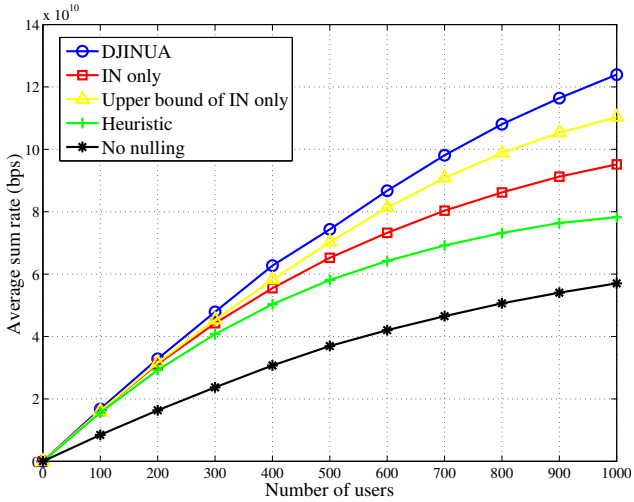

Figure 3.4: Average sum rate versus average number of users. Uniform user distribution, 50 SBS's, $\overline{q} = 6$, 10 antennas at each SBS.

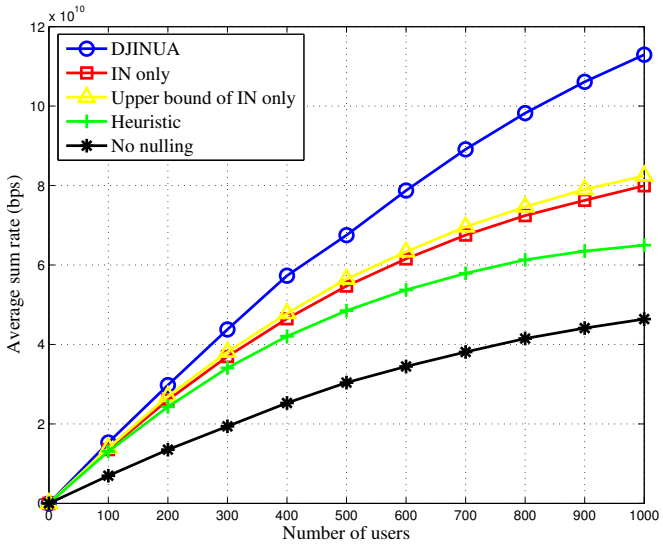

Figure 3.5: Average sum rate versus average number of users. Non-uniform user distribution, 50 SBS's, $\overline{q} = 6$, 10 antennas at each SBS.

The sum rates of different schemes versus the number of SBS's under uniform and non-uniform user distribution are presented in Figs. 3.2 and 3.3, respectively. It can be seen that without interference management, the sum rate first decreases as the SBS's are deployed, since the SINRs of MUEs are significantly reduced. As the number of SBS's continues to grow, the sum rate increases since more users can be served by nearby SBS's. Compared to the No Nulling scheme, a significant performance gain can be achieved by interference nulling, as a result of enhanced SINRs of both MUEs and SUEs. The sum rates with the DJINUA and IN only schemes are higher than the heuristic scheme since both schemes optimize the performance

from the perspective of the entire network. As expected, the DJINUA scheme outperforms the IN only scheme since the user association is jointly considered with interference nulling. We can also observe that the performance of the IN only scheme is close to its upper bound, indicating that the solution with linear approximation is near optimal. The performance gaps between the proposed schemes (DJINUA and IN only) and heuristic scheme become larger as the number of SBS's increases, since the heuristic scheme only achieves a local optimal solution for each BS. The resulting performance loss increases as the network gets larger. Compared to the case of uniform user distribution, the performance gain brought by interference nulling is decreased. This is because the number of users in each SBS varies from 0 to U, thus DoF of each BS is more likely to be under-utilized or insufficient. The performance gap between the DJINUA and IN only is increased compared to the uniform user distribution case, since the DJINUA is adaptive to the change of traffic with efficient use of DoF. Besides, each user and each BS can achieve higher data rate via the propose, compare, and reject operations in the poly matching.

In Figs. 3.4 and 3.5, we compare the sum rates under different numbers of users under uniform and non-uniform user distribution. Due to the same reasons, similar trends are observed for the different schemes. It can be seen when the number of users is sufficiently large, the performances of schemes with interference nulling are increasingly affected by interference, due to the fact that all the DoFs are used. The performance gap between the DJINUA and IN only becomes significant when the number of users is large, showing the importance of DoF-aware joint schedule for user association and interference nulling in case of heavy traffic. The performance gap between the IN only scheme and its upper bound is increased when the number of users becomes large. This is because P becomes smaller as K is increased. Then, the higher-order products are more likely to be 0, and the linear approximation of the upper bound becomes more inaccurate. Similar to Figs. 3.2 and 3.3, the DJINUA scheme is more robust to the variations of traffic compared to other schemes, which also shows the benefits of DoF-aware joint schedule for user association and interference nulling, as well as the benefits of the operations in poly matching.

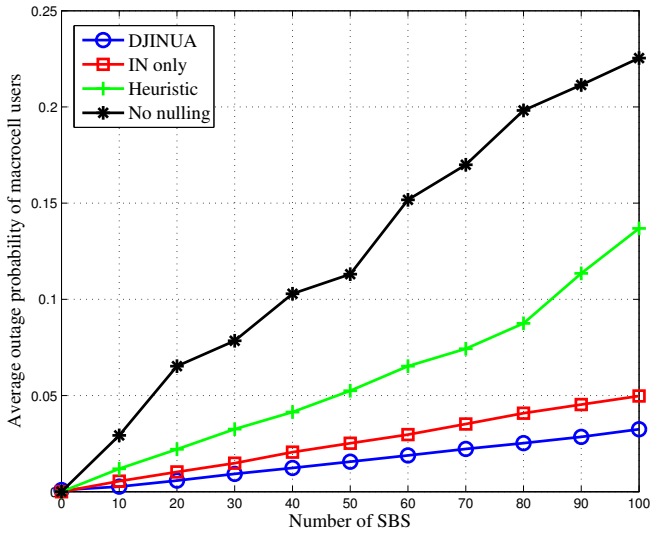

Figure 3.6: Average outage probability of MUs versus number of SBS's. Uniform user distribution, 500 users, $\overline{q} = 6$, 10 antennas at each SBS.

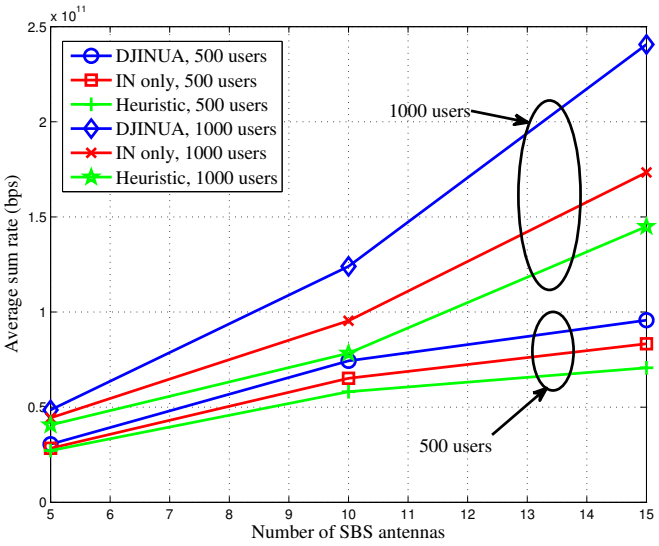

Figure 3.7: Average sum rate versus number of SBS antennas. Uniform user distribution, 50 SBS's, 500 users, $\overline{q} = 6$.

Fig. 3.6 shows the outage performance of macrocell users (MU). We chose to evaluate the MUs since their average SINRs are lower, and hence they are more vulnerable to interference compared to small cell users. Due to the aggregated interference caused by SBS's to MUE, the average outage probability of MUs increases as the number of SBS's grows. It can be seen that with interference nulling performed by SBS's, the average outage probability of MUs is significantly reduced. When the number of SBS's gets large, the outage probabilities of all schemes are increased, since part of the DoFs are used to deal with interference between an

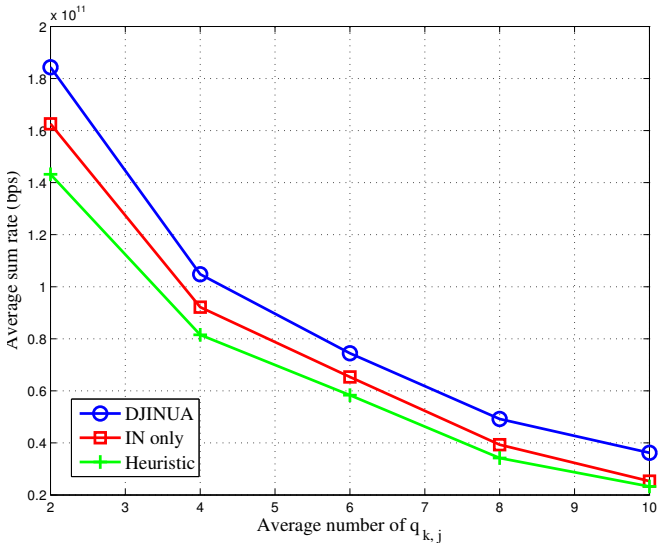

Figure 3.8: Average sum rate versus average number of multipath, $\overline{q}$. Uniform user distribution, 50 SBS's, 500 users, 10 antennas at each SBS.

SBS and SUEs served by other SBS's. The DJINUA scheme achieves the best performance since the users with stronger interference and small number of multipath are more likely to be hold by the BS's for interference nulling. Thus, the DoF of each BS is efficiently used, resulting in most mitigated interference.

The impact of SBS antenna number is evaluated in Fig. 3.7. With increased antenna number at SBS's, more DoF is available for service provision and interference nulling, resulting in improved performance. When the number of users is large, the potential of increased DoF can be fully harnessed, and a near-quadratic performance gain brought by $O\left(N^2\right)$ DoF can be achieved. However, we can also see from the Fig. 3.7 that when the number of users is relatively small, it is unnecessary to increase the number of SBS antennas as no significant performance gain can be achieved. Thus, the SBS antenna configuration should be based on the traffic pattern.

The impact of average number of multipath, $\overline{q}$, is shown in Fig. 3.8. As each user has a larger value of $q_{k,j}$, each BS can put less users into its waiting list, resulting in degraded system performance. In a practical system, if we select a small value of $\overline{q}$ and only consider a certain number of strongest multipath, the evaluation of each users is less accurate, but more users can be included by each BS for service provision or interference nulling. On the other hand, if we select a large value of $\overline{q}$ and take more number of multipath into consideration, a better

information of each user can be obtained, but the less users can be included by each BS for service provision or interference nulling. Thus, such tradeoff should be considered in system design.

3.7 Related Work

DoA information has been considered in recent works to improve the performance of massive MIMO systems [69,70,73,74]. A ESPRIT-based DoA estimation scheme was proposed for 2D massive MIMO systems, and the mean square estimation error was derived. In [70], a multipath channel model was considered, where the channel gain is determined by the steering vector and the attenuation on each path. To reduce the channel estimation complexity and combat the effect of angular spread in DoA-based model, the low-rank property was employed in [73] with a spatial basis expansion model to represent the UL/DL channels. Using the spatial information and CSI of users, the pilot contamination can be mitigated, and the system performance can be enhanced with user scheduling during data transmission period. In a massive MIMO system with two-stage precoding, the angular spread of different user clusters may overlap, resulting in interference. In [74], a graph theory based pattern division scheme was proposed by assigning orthogonal subchannels to overlapping clusters.

Interference management in HetNet is a fundamental challenge, where both inter-tier and intra-tier interference need to be addressed. The major approaches include power control [76], spectrum allocation [**?**], access control [77, 78], beamforming [79], and cognitive radio based interference avoidance [**?**, 80]. Compared to these methods, the interference nulling considered in this chapter is from the perspective of antenna processing, and interference is managed based on the directions of the sources.

3.8 Conclusions

In this chapter, we applied the nested array technology in a massive MIMO HetNet and addressed the problem of joint user association and interference nulling scheduling to maximize the sum rate of MUEs and SUEs. We first considered the case with a given user association,

and formulated the interference nulling scheduling problem as an integer programming problem. We then proposed an approximation solution algorithm, as well as a performance upper bound. We next considered joint user association and interference nulling and proposed a distributed scheme based on a poly matching between users and BS's. The simulation results demonstrated the superior performance of the proposed scheme.

Chapter 4

Enabling Efficient Wireless Backhaul-Based Massive MIMO HetNet with Cross-Layer Design

4.1 Introduction

With the expected massive deployment of small cells in a massive MIMO HetNet, connecting all SBS's to the core network directly with dedicated optical fiber may not be feasible due to significantly increased cost. Alternatively, the SBS's can be connected to the core network by transmitting data to the MBS through backhaul links. In this case, the design of backhaul system is an important issue of a HetNet. Although a massive MIMO HetNet can provide high data rate links between users and BS's, the transmissions between MBS and SBS's may become the bottleneck of the network. Without a reliable backhaul, the aggregated data rates of small cell user equipments (SUE) would be limited by the data rate of the backhaul link. For services with stringent delay requirements, the QoS of users may become unacceptable or even causing outages.

Most existing works have considered wired backhaul between SBS's and MBS, since a wired connection can support high data rate and it is more reliable in general. However, in a HetNet with large number of SBS's, wired connections to each SBS may not be cost-effective or even may be infeasible due to practical constraints. Moreover, the wired backhaul deployment may be highly inefficient when the wireless service provider needs to upgrade or extend the network. Thus, the *wireless backhaul* (WB) has the potential to play an increasingly important role in 5G networks due to its easy and fast deployment, flexibility, and low cost [84–86]. In fact, WB in a massive MIMO HetNet can be quite reliable with proper configurations, especially when massive MIMO are applied with *linear processing* techniques. From the perspective of

an MBS, supporting a WB transmission is equivalent to serving a macrocell user equipment (MUE). With linear processing, e.g., maximum ratio combination (MRC) and maximum ratio transmission (MRT), the reception and precoding are based on linear functions of channel response matrices. When the number of antennas goes to infinity, the inner products of channel vectors of different links grow at a lower rate than that of the number of antennas, the interference between different WBs or MUEs can be averaged out [8]. Thus, the MBS can provide high data rate links to multiple WBs with simple linear processing techniques.

The use of WB in massive MIMO HetNet has drawn some attentions recently [87–90]. In [87], a joint user association and bandwidth allocation scheme was proposed to maximize the downlink sum logarithmic data rate in a massive MIMO HetNet with zero-forcing (ZF) at MBS. A comparison of three WB deployment strategies are presented in [88], namely complete time division duplex, zero division duplex, and zero division duplex with interference rejection. An analytical framework based on stochastic geometry was presented in [89] to study the WB performance in a massive MIMO HetNet with full-duplex small cells, and a closed-form expression of coverage probability was derived. In [90], the network architecture and feasibility issues of WB on the mmWave band were investigated in a dense HetNet with massive MIMO.

Although these works presented several highly efficient approaches, optimal frame design on pilots, i.e., the number of symbols used for pilots in each frame, has not been considered. Here, a frame is defined as a time-frequency resource block and the size of each frame is determined by the coherence time and coherence frequency of all UEs. In each frame, a certain fraction of time is used to transmit symbols that are used as pilots, and these pilots are sent by MUEs and WBs to estimate their channel gains to the MBS. While existing works assume a fixed fraction of time dedicated for pilot, the pilot length, i.e., the number of symbols used for pilots in each frame, can be adaptive to the traffic pattern for performance enhancement. There is clearly a *trade-off* on pilot length here. As discussed, the WBs and MUEs are equivalent from the MBS's point of view. When the pilot length is large, more time is spent on channel estimation at MBS, and a large number of MUEs and WBs can be supported. Moreover, the MUEs and WBs can be allocated with more channels since there is enough time to estimate all these channels. However, as a large fraction of time is dedicated to pilots, the remaining time

for data transmission is small, resulting in a low data rate. When the pilot length is small, the fraction of time for data is increased, but the MUEs and WBs are allocated with less number of channels, which limits the data rates of MUEs and WBs. With a small data rate for WBs, the aggregated data rates of SUEs are limited, resulting in a poor performance.

In this chapter, we investigate the problem of joint frame design, resource allocation, and user association to maximize the downlink sum rate of all users under the WB and fairness constraints. We develop efficient centralized and distributed schemes to obtain the near-optimal solutions to the formulated problem. The main contributions of this chapter are as follows.

- We consider joint pilot length optimization, resource allocation, and user association in a massive MIMO HetNet with WB and linear processing, and provide a rigorous problem formulation.

- We propose a centralized iterative algorithm. The original problem is decomposed into two subproblems and we iteratively solve them until convergence. The first problem is joint pilot length optimization and resource allocation for MUEs and WBs, and we employ a primal decomposition approach to obtain its optimal solution. The second problem is user association, and we obtain its near-optimal solution with a cutting plane approach. An iterative framework is designed to update the parameters of the two subproblems in each iteration to minimize the performance gap between the two problems and guarantee that all constraints are satisfied.

- We propose a distributed scheme by formulating a repeated game among all users, and prove that the game converges to a Nash Equilibrium (NE).

- The performances of the proposed schemes are compared with several benchmark schemes. The simulation results show that performance gains can be as much as more than 100% under certain circumstances.

In the remainder of this chapter, we present the system model and problem formulation in Section 4.2. The centralized and distributed schemes are presented in Sections 4.3 and 4.4, respectively. We discuss our simulation study in Section 4.5. Section 4.7 concludes this chapter.

4.2 Problem Formulation

We consider a noncooperative multi-cell cellular system with focus on a tagged macrocell (denoted as macrocell 0). Macrocell 0 is a two-tier HetNet consisting of an MBS with massive MIMO (indexed by $j = 0$) and J single-antenna SBS's (indexed by $j = 1, 2, \ldots, J$). The payload data of SUEs is transmitted to the core network via WBs between the MBS and SBS's. Then, the reversed time division duplex (RTDD) scheme is a natural choice for the MBS and SBS's [87]. With RTDD, the uplink and downlink transmissions of MBS and SBS's are performed in a reversed pattern, so that an SBS can transmit uplink data to (receive downlink data from) the MBS, and transmit downlink data to (receive uplink data from) SUEs simultaneously. The RTDD scheme is easy to implement in a practical system since it does not require interference cancellation at SBS's. There are K single-antenna mobile users (indexed by $k = 1, 2, \ldots, K$). Each user can be served by either the MBS or an SBS. We define binary variables for user association as

$$x_{k,j} \doteq \begin{cases} 1, & \text{user } k \text{ is associated with BS } j \\ 0, & \text{otherwise,} \end{cases} \quad k = 1, 2, \ldots, K,\ j = 0, 1, \ldots, J. \tag{4.1}$$

The spectrum band owned by the wireless service provider (WSP) is divided into N channels, and the bandwidth of each channel is defined to be the coherence bandwidth of massive MIMO terminals [23]. We assume the MBS adopts *linear processing* schemes with MRC at receiver and MRT at transmitter [8, 28]. From the point of view of MBS, a WB is equivalent to a user to be served. Thus, we can take advantage of the favorable properties of massive MIMO by serving all MUEs and WBs on a same set of channels. This way, they can be put into the beamforming groups on these channels. Due to the law of large numbers, the interference between any two links in a beamforming group can be averaged out. From the perspective of an SBS, a WB is also equivalent to a user to be served. However, since the SBS's are assumed to be equipped with single antenna, they cannot perform interference mitigation in the spatial domain or self-interference cancelation. Hence, orthogonal resources must be assigned between WBs and SUEs to avoid mutual interference. Consequently, we assume that a proportion of

α of the whole bandwidth is allocated to WBs and MUEs, and the rest $(1-\alpha)$ is allocated to SUEs. Note that α needs to be consistent across all macrocells to avoid cross-tier interference between cell-edge users, and it is predetermined by the wireless service provider. The frame structure considered in this chapter is shown in Fig. 4.1.

We assume both the bandwidth of each frame and the bandwidth of a channel equal to the coherence bandwidth of all MUEs and WBs, given as W_c. Then, each frame corresponds to a specific interval on a channel. The duration of a frame is T_c seconds, which equals to the coherence time of all MUEs and WBs. Thus, the channel gains are constant in a frame and each frame can be viewed as a *coherence block*. The interval of a symbol is T_s seconds, which consists of T_u seconds for useful symbols and $T_g = T_s - T_u$ seconds for guard interval. Let Δ_f be the spacing of subcarriers, then T_u is given as $T_u = 1/\Delta_f$. Within a coherence bandwidth, there are W_c/Δ_f subcarriers. Hence, the channel response is constant over $N_{\mathrm{sm}} = W_c/\Delta_f$ consecutive subcarriers in each symbol. Let τ be the pilot length, i.e., the number of OFDM symbols dedicated for pilots in each frame. Then, the number of terminals that can be supported in each frame is τN_{sm}. Therefore, the total number of MUEs and WBs that can be served by the MBS on each channel within the interval of a frame is upper bounded by τN_{sm}.

Given the available spectrum band for MUEs and WBs, we define the following resource allocation indicators

$$a_{k,n} \doteq \begin{cases} 1, & \text{channel } n \text{ is allocated to MUE } k \\ 0, & \text{otherwise,} \end{cases}$$
$$k = 1, 2, \ldots, K,\ n = 1, \ldots, \alpha N. \tag{4.2}$$

$$b_{j,n} \doteq \begin{cases} 1, & \text{channel } n \text{ is allocated to SBS } j\text{'s WB} \\ 0, & \text{otherwise,} \end{cases}$$
$$j = 1, 2, \ldots, J,\ n = 1, \ldots, \alpha N. \tag{4.3}$$

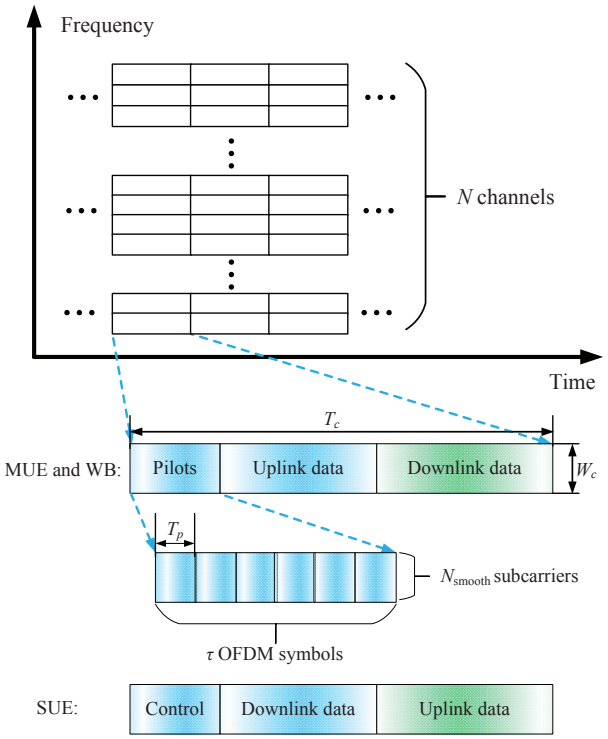

Figure 4.1: Resource allocation and frame structure of a massive MIMO HetNet with wireless backhaul.

According to our analysis, we have

$$\sum_{k=1}^{K} a_{k,n} + \sum_{j=1}^{J} b_{j,n} \leq \tau N_{\text{sm}}, \ n = 1, \ldots, \alpha N. \tag{4.4}$$

In a massive MIMO system, the effects of fast fading and noise vanish as the number of antennas goes to infinity; the only possible interference comes from the UEs that share the same pilot sequence [8]. Thus, the only factor that limits the performance of a massive MIMO system with linear processing is pilot contamination. For user k connecting to the MBS in macrocell 0, let macrocell l be the neighboring macrocell(s) that uses the same pilot sequence as user k. The downlink signal to interference ratio (SIR) of user k when it connects to the MBS in the tagged macrocell, $\gamma_{k,0}$, is

$$\gamma_{k,0} = \beta_{k,0}^2 / \sum_{l \neq 0} \beta_{k,l}^2, \tag{4.5}$$

where $\beta_{k,0}$ is the factor accounting for the propagation loss and shadowing effects between the MBS and user k, and $\beta_{k,l}$ accounts for the propagation loss and shadowing factor between user k and the MBS in macrocell l. When neighboring macrocells use different values of τ, an MBS receives not only the pilot signals of users from other cells, but also uplink data signals from other cells. As analyzed in [28], the non-orthogonal uplink data signals also contaminate the channel estimation of other cells, and the resulting interference is a random variable bounded by the interference caused by pilot signals. Hence, we use (4.5) as a worst-case approximation in case the SIR cannot be measured by the MBS due to technical limits. When the values of τ are close to each other in different macrocells, such approximation would be highly reliable. Due to the mobility of users, we assume that $\gamma_{k,0}$ is updated with a period of T seconds.

The data rate of user k is given by [8]

$$R_{k,0} = \sum_{n=1}^{\alpha N} a_{k,n} \left(1 - \frac{T_p}{T_c}\tau\right) \left(\frac{T_u}{T_s}\right) \log\left(1 + \gamma_{k,0}\right), \tag{4.6}$$

where T_p is the time spent to transmit pilot for one user and $T_p = T_s$.[1] Due to channel reciprocity of the TDD mode, the CSI is acquired by the MBS using uplink pilots. Then, $\gamma_{k,0}$ and $R_{k,0}$ can be obtained by the MBS.

Similarly, let γ_j be the downlink SIR of WB between the MBS and SBS j, it is given by

$$\gamma_j = \beta_{j,0}^2 / \sum_{l \neq 0} \beta_{j,l}^2, \tag{4.7}$$

where $\beta_{j,0}$ is the factor accounts for the propagation loss and shadowing effects between the MBS and SBS j, and $\beta_{j,l}$ is the propagation loss and shadowing factor between SBS j and the MBS in macrocell l.

[1]Note that, the value of T_p can also be optimized based physical layer analysis. According to (4.6), a small value of T_p reduces the channel estimation overhead and increases $R_{k,0}$. However, the channel estimation quality may be degraded, resulting in decreased $R_{k,0}$. Due to space limit, we focus on frame level analysis and network scheduling problems, the potential of optimizing T_p with physical layer analysis can be investigated in future work.

The data rate of the WB for SBS j is then given as

$$C_j = \sum_{n=1}^{\alpha N} b_{j,n} \left(1 - \frac{T_p}{T_c}\tau\right) \left(\frac{T_u}{T_s}\right) \log\left(1 + \gamma_j\right). \tag{4.8}$$

We assume that the time interval for uplink pilots of MUEs and WBs are used to send control information from SBS's to SUEs, including CSI, power and channel schedule of SUEs. We also assume that equal resource allocation is applied to SUEs served by the same SBS so that proportional fairness can be achieved [87]. Let $\gamma_{k,j}$ be the average signal to noise plus interference ratio (SINR) of user k connecting to SBS j over a time period. The achievable data rate is given as

$$R_{k,j} = \left(1 - \frac{T_p}{T_c}\tau\right) \left(\frac{T_u}{T_s}\right) \frac{(1-\alpha)N}{\sum_{k=1}^{K} x_{k,j}} \log\left(1 + \gamma_{k,j}\right). \tag{4.9}$$

We assume that the powers of SBS's and SUEs are adjusted to proper values so that the interference between different small cell users are controlled at an acceptable level. Unlike the MBS with massive MIMO, the effect of fast fading exists on the channel between an SUE and an SBS, resulting in frequently varying CSI. Therefore, it is infeasible to use the instantaneous CSI for scheduling purposes. To this end, $\gamma_{k,j}$ is based on the *time-averaged* CSI measured by the SBS over T seconds in the previous period, and it is updated every T seconds.

We aim to maximize the sum rate of a massive MIMO HetNet. Let $\mathbf{x}$, $\mathbf{a}$, and $\mathbf{b}$ denote the matrices of $\{x_{k,j}\}$, $\{a_{k,n}\}$, and $\{b_{j,n}\}$, respectively. The problem is formulated as

$$\mathbf{P1}: \max_{\{\mathbf{x},\mathbf{a},\mathbf{b},\tau\}} \sum_{k=1}^{K}\sum_{j=0}^{J} x_{k,j} R_{k,j} \tag{4.10}$$

subject to:

$$\sum_{j=0}^{J} x_{k,j} \leq 1,\ k = 1, 2, \ldots, K \tag{4.11}$$

$$\sum_{k=1}^{K} x_{k,j} \leq S_j,\ j = 0, 1, \ldots, J \tag{4.12}$$

$$\sum_{k=1}^{K} a_{k,n} + \sum_{j=1}^{J} b_{j,n} \leq \tau N_{\mathrm{sm}},\ n = 1, \ldots, \alpha N \tag{4.13}$$

$$\sum_{n=1}^{\alpha N} a_{k,n} \leq E_k,\ k = 1, 2, \ldots, K \tag{4.14}$$

$$\sum_{n=1}^{\alpha N} b_{j,n} \leq F_j,\ j = 1, 2, \ldots, J \tag{4.15}$$

$$\sum_{k=1}^{K} x_{k,j} R_{k,j} \leq C_j,\ j = 1, 2, \ldots, J \tag{4.16}$$

$$\tau \leq \tau_{\max},\ \tau \in \mathcal{N}^+ \tag{4.17}$$

$$a_{k,n} \in \{0, 1\},\ b_{j,n} \in \{0, 1\},\ x_{k,j} \in \{0, 1\},$$
$$n = 1, \ldots, \alpha N, k = 1, \ldots, K, j = 0, \ldots, J. \tag{4.18}$$

In problem **P1**, constraint (4.11) is because each user can connect to at most one BS. We enforce an upper bound on the number of users that can be served by each BS in (4.12) to guarantee the QoS of users. Constraint (4.13) is directly from (4.4). By enforcing an upper bound on the number of channels that can be accessed by user k, constraint (4.14) is to guarantee fairness among the MUEs. Without such constraint, MUEs with high SIRs would be allocated with more channels than those with low SIRs, resulting in poor fairness performance.[2] Thus, the value of E_k for an MUE with high SIR is set to be lower than an MUE with low SIR.[3] Similarly, constraint (4.15) is to guarantee fairness among the WBs. Constraint (4.16) is due to the fact that the data rate of WB for SBS j should be larger than or equal to the sum rate of all SUEs served by SBS j. Constraint (4.17) enforces an upper bound for the number of symbols that are allocated to pilot transmissions. Since we assume both $\gamma_{k,0}$ and $\gamma_{k,j}$ are updated with the period of T, problem **P1** is also solved with the period of T.

[2]Due to the channel hardening effect of a massive MIMO system, the channel gains across different frequencies are close to each other [95]. Thus, the dominant factor that impacts the performance of WBs and MUEs is the number of allocated channels. However, in other application scenarios where the channel response varies significantly over different frequencies, e.g., in a mmWave network, frequency domain scheduling should be considered. Some existing approaches can be applied include proportional fairness scheduling [91], bipartite matching based algorithm [81].

[3]The proper values of E_k and F_j depend on network topology, traffic pattern, and QoS requirement of users. In a specific system, E_k and F_j can be dynamically adjusted based on the QoS of users. When the data rate of a MUE or WB at the edge of cell is lower than a threshold, the values of E_k and F_j for the MUEs and WBs with highest data rates will be lowered in the next period. The adjustment strategy of E_k and F_j can be done with an offline training process for each cell.

4.3 Centralized Solution Algorithm

In this section, we develop a centralized iterative scheme to obtain the near optimal solution of **P1**. Problem **P1** is an integer programming problem with both coupling variables and coupling constraints, and constraint (4.16) is a nonlinear coupling constraint of two sets of variables. Thus, standard optimization techniques cannot be directly applied for the optimal solution.

To make the problem tractable, we decompose problem **P1** into (i) *WB and MUE resource allocation and pilot length optimization* problem and (ii) *user association* problem, and iteratively solve the two problems until convergence. At each iteration, we update the constraints of each problem to guarantee that all constraints of the original problem are satisfied.

4.3.1 Resource Allocation and Pilot Optimization

As can be seen in (4.6) and (4.9), $R_{k,0}$ is determined by $\mathbf{a}$; and $R_{k,j}$, $j = 1, ..., J$, is limited by $\mathbf{b}$. Due to constraint (4.16), the sum rate of all MUEs and WBs naturally serves as an upper bound for the sum rate of all users. Thus, it is reasonable to try to maximize this upper bound and iteratively tighten the gap, so that the final solution is a close approximation for the optimal solution of Problem **P1**. The problem of maximizing the sum rate of all MUEs and WBs for a given $\mathbf{x}$ is presented as follows.

$$\begin{aligned}\mathbf{P2}: \max_{\{\mathbf{a},\mathbf{b},\tau\}} & \left(1-\frac{T_p}{T_c}\tau\right)\cdot \\ & \left\{\sum_{k=1}^{K}\sum_{n=1}^{\alpha N} a_{k,n}\log\left(1+\gamma_{k,0}\right)+\sum_{j=1}^{J}\sum_{n=1}^{\alpha N} b_{j,n}\log\left(1+\gamma_j\right)\right\} \\ \text{subject to: } & (4.13)-(4.18).\end{aligned} \tag{4.19}$$

Note that, constraint (4.16) can be written as $\sum_{n=1}^{\alpha N} b_{j,n} \geq \frac{\sum_{k=1}^{K} x_{k,j} R_{k,j}}{\log(1+\gamma_j)}$. Since $\sum_{n=1}^{\alpha N} b_{j,n}$ is always an integer, (4.16) is equivalent to $\sum_{n=1}^{\alpha N} b_{j,n} \geq \left\lceil \frac{\sum_{k=1}^{K} x_{k,j} R_{k,j}}{\log(1+\gamma_j)} \right\rceil$.

Suppose constraint (4.16) has already been satisfied for the WB of SBS j, then allocating more resources to this WB can not improve the actual sum rate of the users served by SBS j, while it potentially increases the value of τ, resulting in degraded system performance. Thus,

(4.16) is an active constraint in problem **P2**. We have $\sum_{n=1}^{\alpha N} b_{j,n} = \left\lceil \frac{\sum_{k=1}^{K} x_{k,j} R_{k,j}}{\log(1+\gamma_j)} \right\rceil$. Combining this constraint with (4.15), we have

$$\sum_{n=1}^{\alpha N} b_{j,n} = \min \left\{ \left\lceil \frac{\sum_{k=1}^{K} x_{k,j} R_{k,j}}{\log\left(1+\gamma_j\right)} \right\rceil, F_j \right\} .j = 1, 2, \ldots, J. \tag{4.20}$$

To solve problem **P2**, we first relax the integer constraints of $\mathbf{a}$, $\mathbf{b}$, and τ by allowing them to take any values in $[0, 1]$.

Lemma 18 *The relaxed problem of* **P2***,* **P2-Relaxed***, is a convex optimization problem.*

Proof: The objective function of **P2-Relaxed** is a sum of quadratic terms and linear functions, with the quadratic terms given as $-\tau a_{k,n}$ and $-\tau b_{j,n}$. It can be easily verified that the Hessian matrices of such quadratic terms are negative definite. Thus, the objective function is concave. Since all constraints are linear, **P2-Relaxed** is a convex optimization problem. ■

Since the decision variables are coupled in the constraints, we use a primal decomposition to transform problem **P2-Relaxed** into two levels of problems [94]. At the lower level, we find optimal solution of $\mathbf{a}$ and $\mathbf{b}$ for a given τ. Based on the solution of the lower level problem, the optimal value of τ is then obtained with a subgradient approach.

Optimal Solution of $\mathbf{a}$ and $\mathbf{b}$ for Given τ

With given τ, we have the following lower level problem of **P2-Relaxed**.

$$\begin{aligned} \mathbf{P3}: \max_{\{\mathbf{a},\mathbf{b}\}} & \sum_{k=1}^{K} \sum_{n=1}^{\alpha N} a_{k,n} \log\left(1+\gamma_{k,0}\right) \\ & + \sum_{j=1}^{J} \sum_{n=1}^{\alpha N} b_{j,n} \log\left(1+\gamma_j\right) \\ \text{subject to: } & (4.13), (4.14), (4.18), \text{and } (4.20). \end{aligned} \tag{4.21}$$

We can see that **P3** is a linear programming (LP), which can be solved with efficient methods such as simplex method. To analyze its property, we transform **P3** into the standard form by

concatenating the columns of $\mathbf{a}$ and $\mathbf{b}$ alternately, given as

$$\tilde{\mathbf{y}} = [a_{1,1}, \ldots, a_{K,1}, b_{1,1}, \ldots, b_{J,1}, a_{1,2}, \ldots, a_{K,2}, \\ b_{1,2}, \ldots, b_{J,2}, \ldots, a_{1,\alpha N}, \ldots, a_{K,\alpha N}, b_{1,\alpha N}, \ldots, b_{J,\alpha N}]^T. \tag{4.22}$$

Let $\mathbf{Z}$ be the constraint matrix corresponding to $\tilde{\mathbf{y}}$, as

$$\mathbf{Z} \doteq \begin{pmatrix} 1\,1 \cdots 1 & 0\,0 \cdots 0 & \cdots & 0\,0 \cdots 0 \\ 0\,0 \cdots 0 & 1\,1 \cdots 1 & \cdots & 0\,0 \cdots 0 \\ \vdots & \vdots & & \vdots \\ 0\,0 \cdots 0 & 0\,0 \cdots 0 & \cdots & 1\,1 \cdots 1 \\ 1\,0 \cdots 0 & 1\,0 \cdots 0 & \cdots & 1\,0 \cdots 0 \\ 0\,1 \cdots 0 & 0\,1 \cdots 0 & \cdots & 0\,1 \cdots 0 \\ \ddots & \ddots & \vdots & \ddots \\ 0\,0 \cdots 1 & 0\,0 \cdots 1 & \cdots & 0\,0 \cdots 1 \end{pmatrix} \tag{4.23}$$

The right hand side (RHS) of the LP is a $(\alpha N + J + K) \times 1$ vector, given by

$$\mathbf{d} = [\tau N_{\text{sm}}, ..., \tau N_{\text{sm}}, E_1, ..., E_K, \theta_1, ..., \theta_J]^T, \tag{4.24}$$

where $\theta_j = \min\left\{\left\lceil \sum_{k=1}^{K} x_{k,j} R_{k,j} / \log\left(1+\gamma_j\right)\right\rceil, F_j\right\}$.

Lemma 19 *The constraint matrix* $\mathbf{Z}$ *is totally unimodular.*

Proof: The proof can be found in [4]. ■

Property 1 *If the constraint matrix of an LP satisfies totally unimodularity, and the RHS is integral, then it has all integral vertex solutions [44].*

Property 2 *If an LP has feasible optimal solutions, then at least one of the feasible optimal solutions occurs at a vertex of the polyhedron defined by its constraints [45].*

Lemma 20 *All the decision variables in the optimal solution to the relaxed LP, problem* **P3***, are integers in* $\{0, 1\}$.

Proof: This lemma directly follows Lemma 19, Property 1, and Property 2. ■

Optimal Value of τ

Denote $g\left(\mathbf{a}\left(\tau\right),\mathbf{b}\left(\tau\right),\tau\right)$ and $f\left(\mathbf{a}\left(\tau\right),\mathbf{b}\left(\tau\right)\right)$ as the values of objective functions of **P2-Relaxed** and **P3** for a given τ, which are given in (4.19) and (4.21), respectively. Let $g^{*}\left(\tau\right)$ and $f^{*}\left(\tau\right)$ be their optimal values for a given τ, respectively. At the higher level of problem **P2-Relaxed**, we find the optimal value of τ by solving the following problem.

$$\mathbf{P4}: \max_{\{\tau\}} g^{*}(\tau). \tag{4.25}$$

Consider the objective function of **P2-Relaxed**, given as

$$g\left(\mathbf{a}\left(\tau\right),\mathbf{b}\left(\tau\right),\tau\right)=\left(1-\frac{T_p}{T_c}\tau\right)\left(f\left(\mathbf{a}\left(\tau\right),\mathbf{b}\left(\tau\right)\right)\right). \tag{4.26}$$

Maximizing (4.26) is equivalent to maximizing the following

$$\log\left(1-\frac{T_p}{T_c}\tau\right)+\log\left[f\left(\mathbf{a}\left(\tau\right),\mathbf{b}\left(\tau\right)\right)\right]. \tag{4.27}$$

Hence, problem **P4** is equivalent to the following problem

$$\max_{\{\tau\}}\left\{\log\left(1-\frac{T_p}{T_c}\tau\right)+\log\left[f\left(\mathbf{a}^{*}\left(\tau\right),\mathbf{b}^{*}\left(\tau\right)\right)\right]\right\} \tag{4.28}$$
$$\text{subject to: } (4.17).$$

Let $h_1\left(\tau\right)=\log\left(1-\frac{T_p}{T_c}\tau\right)$, $h_2\left(\tau\right)=\log\left[f\left(\mathbf{a}^{*}\left(\tau\right),\mathbf{b}^{*}\left(\tau\right)\right)\right]$, and $h\left(\tau\right)=h_1\left(\tau\right)+h_2\left(\tau\right)$. Since **P2-Relaxed** is a convex problem according to Lemma 18, we can apply primal decomposition to optimize $h_1\left(\tau\right)$ and $h_2\left(\tau\right)$ separately [94]. It can be easily verified that $h_1\left(\tau\right)$ is a differentiable concave function. For any τ and τ', we have

$$\log\left(1-\frac{T_p}{T_c}\tau\right)\leq\log\left(1-\frac{T_p}{T_c}\tau'\right)-\frac{T_p}{T_c-T_p\tau'}\left(\tau-\tau'\right).$$

Then, τ can be updated with the following gradient approach to maximize $h_1(\tau)$.

$$\tau^{[t+1]} = \tau^{[t]} - \frac{T_p}{T_c - T_p\tau^{[t]}}\rho^{[t]}, \tag{4.29}$$

where t is the index of iteration and $\rho^{[t]}$ is the step size.

To obtain the optimal solution of $h_2(\tau)$, we consider the following optimization problem

$$\mathbf{P5}: \max_{\{\mathbf{a},\mathbf{b}\}} \log\left[f\left(\mathbf{a}(\tau), \mathbf{b}(\tau)\right)\right]$$

$$\text{subject to: } (4.13), (4.14), (4.18), \text{and } (4.20).$$

Lemma 21 *Strong duality holds for problem* **P5**.

Proof: Since problem **P5** is a convex problem, all the constraints are linear and the Slater condition reduces to feasibility [41, 87]. Thus strong duality holds. ■

Let λ_n^* be the optimal value of Lagrangian multiplier corresponding to the constraint $\sum_{k=1}^{K} a_{k,n} + \sum_{j=1}^{J} b_{j,n} \leq \tau N_{sm}$. We consider the optimal solutions to **P5** for two different values, τ' and τ. Then, we have

$$\begin{aligned}
&h_2(\tau') = \log\left[f\left(\mathbf{a}^*(\tau'), \mathbf{b}^*(\tau')\right)\right] \\
&\overset{(a)}{=} \mathcal{L}\left(\mathbf{a}^*(\tau'), \mathbf{b}^*(\tau'), \boldsymbol{\lambda}^*(\tau'), \boldsymbol{\mu}^*(\tau'), \boldsymbol{\nu}^*(\tau'), \boldsymbol{\eta}^*(\tau')\right) \\
&\overset{(b)}{\geq} \mathcal{L}\left(\mathbf{a}^*(\tau), \mathbf{b}^*(\tau), \boldsymbol{\lambda}^*(\tau'), \boldsymbol{\mu}^*(\tau'), \boldsymbol{\nu}^*(\tau'), \boldsymbol{\eta}^*(\tau')\right) \\
&= \log\left[f\left(\mathbf{a}^*(\tau), \mathbf{b}^*(\tau)\right)\right] + \sum_{n=1}^{\alpha N} \lambda_n^*(\tau')\left(\tau' N_{sm} - \delta_n^*(\tau)\right) + \Phi \\
&= \log\left[f\left(\mathbf{a}^*(\tau), \mathbf{b}^*(\tau)\right)\right] + \sum_{n=1}^{\alpha N} \lambda_n^*(\tau')\left(\tau N_{sm} - \delta_n^*(\tau)\right) \\
&+ \Phi + \sum_{n=1}^{\alpha N} \lambda_n^*(\tau')\left(\tau' N_{sm} - \tau N_{sm}\right) \\
&\overset{(c)}{\geq} h_2(\tau) + N_{sm} \sum_{n=1}^{\alpha N} \lambda_n^*(\tau')\left(\tau' - \tau\right),
\end{aligned} \tag{4.30}$$

where $\delta_n^*(\tau) = \sum_{k=1}^{K} a_{k,n}^*(\tau) + \sum_{j=1}^{J} b_{j,n}^*(\tau)$, $\boldsymbol{\mu}$, $\boldsymbol{\nu}$, and $\boldsymbol{\eta}$ are the Lagrangian multipliers corresponding to other constraints. Φ is given as

$$\Phi = \sum_{k=1}^{K} \mu_k^*(\tau') \left(E_k - \sum_{n=1}^{\alpha N} a_{k,n}(\tau) \right) + \sum_{j=1}^{J} \nu_j^*(\tau') \left(F_j - \sum_{n=1}^{\alpha N} b_{j,n}(\tau) \right) + \sum_{j=1}^{J} \eta_j^*(\tau') \left(\sum_{n=1}^{\alpha N} b_{j,n}(\tau) - \left\lceil \frac{R_{k,j}}{\log(1+\gamma_j)} \right\rceil \right). \tag{4.31}$$

In (4.30), equality (a) is due to strong duality, inequality (b) is due to the optimality of $\mathbf{a}^*(\tau')$ and $\mathbf{b}^*(\tau')$, and inequality (c) is due to the constraints of problem **P5** and the nonnegativity of all Lagrangian multipliers.

It follows (4.30) that

$$h_2(\tau) \leq h_2(\tau') + N_{sm} \sum_{n=1}^{\alpha N} \lambda_n^*(\tau')(\tau - \tau'). \tag{4.32}$$

By definition, $N_{sm} \sum_{n=1}^{\alpha N} \lambda_n^*(\tau)$ is a subgradient of $h_2(\tau)$. The maximum value of $h_2(\tau)$ can be obtained by

$$\tau^{[t+1]} = \tau^{[t]} + N_{sm} \sum_{n=1}^{\alpha N} \lambda_n^{*[t]} \rho^{[t]} \tag{4.33}$$

Lemma 22 *Problem* **P4** *can be solved by the following subgradient method.*

$$\tau^{[t+1]} = \tau^{[t]} + \left(N_{sm} \sum_{n=1}^{\alpha N} \lambda_n^{*[t]} - \frac{T_p}{T_c - T_p \tau^{[t]}} \right) \rho^{[t]}. \tag{4.34}$$

Proof: According the principles of primal decomposition, $N_{sm} \sum_{n=1}^{\alpha N} \lambda_n^{*[t]} - \frac{T_p}{T_c - T_p \tau^{[t]}}$ is a subgradient of $h(\tau)$, τ can be updated by combining (4.29) and (4.33). The optimal value of τ can be achieved until iteration converges. ■

There is a nice interpretation for (4.34). In each update, $N_{\mathrm{sm}} \sum_{n=1}^{\alpha N} \lambda_n^{*[t]}$ indicates the performance gain obtained by allocating more pilot symbols to WBs and MUEs, i.e., to increase

τ. The second part, $\frac{T_p}{T_c - T_p\tau^{[t]}}$ indicates the performance loss caused by the reduced number of data symbols.

Denote $\eta^{[t]}$ as the subgradient of $h(\tau)$, $\eta^{[t]} = N_{sm}\sum_{n=1}^{\alpha N}\lambda_n^{*[t]} - \frac{T_p}{T_c - T_p\tau^{[t]}}$, the convergence of the τ is shown in the following lemma.

Lemma 23 *With step size set as* $\rho^{[t]} = \frac{h(\tau^*) - h(\tau^{[t]})}{(\eta^{[t]})^2}$*, the sequence* $h(\tau^{[t]})$ *converges to its optimal value* $h(\tau^*)$ *with a speed faster than* $\{1/\sqrt{t}\}$ *as* $t \to \infty$*.*

Proof: Consider the optimality gap of τ, we have

$$
\begin{aligned}
&(\tau^{[t+1]} - \tau^*)^2 \le \left(\tau^{[t]} + \frac{h(\tau^*) - h(\tau^{[t]})}{(\eta^{[t]})^2}\eta^{[t]} - \tau^*\right)^2 \\
&= (\tau^{[t]} - \tau^*)^2 + \frac{\left(h(\tau^*) - h(\tau^{[t]})\right)^2}{(\eta^{[t]})^2} + 2(\tau^{[t]} - \tau^*)\eta^{[t]}\frac{h(\tau^*) - h(\tau^{[t]})}{(\eta^{[t]})^2} \\
&\le (\tau^{[t]} - \tau^*)^2 - \frac{\left(h(\tau^*) - h(\tau^{[t]})\right)^2}{(\eta^{[t]})^2} \le (\tau^{[t]} - \tau^*)^2 - \frac{\left(h(\tau^*) - h(\tau^{[t]})\right)^2}{\widehat{\eta}^2},
\end{aligned}
$$

where $\widehat{\eta}$ is an upper bound of $|\eta^{[t]}|$. The first inequality is because $\tau^{[t+1]}$ should project to $[0, \tau_{\max}]$, the second inequality is due to the property of subgradient, given as $(\tau^{[t]} - \tau^*)\eta^{[t]} \le h(\tau^{[t]}) - h(\tau^*)$. Summing the above inequality from $t = 1$ to $t \to \infty$, we have

$$
\sum_{t=1}^{\infty}\left(h(\tau^*) - h(\tau^{[t]})\right)^2 \le \widehat{\eta}^2(\tau^{[1]} - \tau^*)^2. \tag{4.35}
$$

Suppose for contradiction, $\lim_{t\to\infty}\left(h(\tau^*) - h(\tau^{[t]})\right)\sqrt{t} > 0$. Then, there must be a sufficiently large t' and a positive number ξ such that $\left(h(\tau^*) - h(\tau^{[t]})\right)\sqrt{t} > \xi, \forall t \ge t'$. Taking the square sum from t' to ∞, we have

$$
\sum_{t=t'}^{\infty}\left(h(\tau^*) - h(\tau^{[t]})\right)^2 \ge \xi^2\sum_{t=t'}^{\infty}\frac{1}{t} = \infty. \tag{4.36}
$$

It can seen that (4.36) contradicts (4.35). Thus, the hypothesis does not hold, we have

$$
\lim_{t\to\infty}\frac{h(\tau^*) - h(\tau^{[t]})}{1/\sqrt{t}} = 0, \tag{4.37}
$$

Algorithm 4: WB and MUE Resource Allocation and Pilot Length Optimization

Initialize τ ;
do
  Solve problem **P5** to obtain $\lambda_n^*(\tau)$;
  Update τ with (4.34) ;
while (τ *does not converge and* $\tau \leq \tau_{\max}$);
if $\tau < \tau_{\max}$ **then**
  Solve **P3** with $\lfloor \tau^* \rfloor$ and $\lceil \tau^* \rceil$ to obtain $\mathbf{a}(\lceil \tau^* \rceil)$, $\mathbf{b}(\lceil \tau^* \rceil)$, $\mathbf{a}(\lfloor \tau^* \rfloor)$, and $\mathbf{b}(\lfloor \tau^* \rfloor)$;
  Use the results to compare the values of objective functions of **P2**. Then, $\tau^* = \arg\max_{\{\lfloor \tau^* \rfloor, \lceil \tau^* \rceil\}} \{g^*(\lfloor \tau^* \rfloor), g^*(\lceil \tau^* \rceil)\}$;
else
  Set $\tau^* = \tau_{\max}$;
end
Use τ^* to solve problem **P3**, and obtain the optimal $\mathbf{a}$ and $\mathbf{b}$;

this indicates that $h(\tau^{[t]})$ converges with a speed faster than that of $1/\sqrt{t}$. ■

Note that, the optimal τ to **P2-Relaxed** may not be an integer. Since **P2-Relaxed** is a convex problem, a simple way to find the optimal τ to **P2** is to compare the objective values of problem **P2** under $\lfloor \tau^* \rfloor$ and $\lceil \tau^* \rceil$, and select the larger one. As discussed in Lemma 20, the optimal solution to **P2-Relaxed** are integers for any given integer value of τ. Thus, such solution is also optimal to **P2**, we conclude that the optimal solution of **P2** can be obtained.

The procedure of the proposed WB and MUE resource allocation and pilot length optimization scheme is summarized in Algorithm 4.

Lemma 24 *The complexity of Algorithm 4 is upper bounded by* $1/\varepsilon_1^2\varepsilon_2^2$*, where* ε_1 *is the threshold of convergence for* τ*,* ε_2 *is the threshold of convergence for* $\boldsymbol{\lambda}$*.*

Proof: According to Lemma 23 and (4.37), for a sufficiently large t and a sufficiently small ε_1, we have $h(\tau^*) - h(\tau^{[t]}) < 1/\sqrt{t}$. Thus, when $1/\sqrt{t} > \varepsilon_1$, $h(\tau^*) - h(\tau^{[t]})$ is guaranteed to be smaller than ε_1. Consequently, it takes less than $1/\varepsilon_1^2$ steps for the sequence $h(\tau^{[t]})$ to achieve an optimality gap that is less than ε_1, $t < 1/\varepsilon_1^2$. In the same way, the number of iterations for the convergence of $\boldsymbol{\lambda}$ is upper bounded by $1/\varepsilon_2^2$.

In Algorithm 4, each update of τ requires a set of optimal $\boldsymbol{\lambda}$. Thus, the total number of variable updates is upper bounded by $1/\varepsilon_1^2\varepsilon_2^2$, the complexity of Algorithm 4 is upper bounded by $1/\varepsilon_1^2\varepsilon_2^2$. ■

4.3.2 User Association under WB Constraints

For a given set of $\mathbf{a}$, $\mathbf{b}$, and τ, **P1** is reduced to the following user association problem.

$$\mathbf{P6}: \max_{\{\mathbf{x}\}} \sum_{k=1}^{K} \sum_{j=0}^{J} x_{k,j} R_{k,j} \tag{4.38}$$

subject to: $(4.11), (4.12),$ and (4.16)

$$x_{k,j} \in \{0,1\}, \ k = 1, 2, \ldots, K, \ j = 0, 1, \ldots, J.$$

Constraint (4.16) can be rewritten as

$$\sum_{k=1}^{K} x_{k,j} \left(\log\left(1+\gamma_{k,j}\right) - \frac{\sum_{n=1}^{\alpha N} b_{j,n} \log\left(1+\gamma_j\right)}{(1-\alpha) N} \right) \leq 0,$$
$$j = 1, 2, \ldots, J, \tag{4.39}$$

which is a linear constraint on $\mathbf{x}$.

To solve **P6**, we first relax the integer constraint of $\mathbf{x}$ by allowing all $x_{k,j}$ to take any value between $[0, 1]$. Denote the relaxed problem as **P6-Relaxed**. The objective function of **P6-Relaxed** includes a weighted sum of $\frac{\sum_{k=1}^{K} x_{k,j} \log\left(1+\gamma_{k,j}\right)}{\sum_{k=1}^{K} x_{k,j}}$, which is non-convex. Thus, only local optimal solution can be achieved with standard optimization techniques. However, if the values of $Q_j = \sum_{k=1}^{K} x_{k,j}$ are given, **P6-Relaxed** reduces to an LP.

Since $Q_j \leq S_j$, the optimal solution of **P6-Relaxed** can be obtained by searching all possible combinations of $\mathbf{Q} = \{Q_1, ..., Q_J\}$ and solve the corresponding LPs. However, this results in a high complexity as a number of $\prod_{j=1}^{J} S_j$ LPs need to be solved. Therefore, we use this approach to obtain the initial optimal values of $\mathbf{Q}$ and update it with a more efficient approach. Recall that the system states are updated every T. Thus, in a low mobility environment, we can make use of $\mathbf{Q}$ in the previous period as an approximation to the $\mathbf{Q}$ of the current period. Then, $\{R_{k,j}\}$ becomes independent of $\mathbf{x}$, given as

$$R_{k,j} = \frac{1}{Q_j} \left(1 - \frac{T_p}{T_c}\tau\right) \left(\frac{T_u}{T_s}\right) (1-\alpha) N \log\left(1+\gamma_{k,j}\right).$$

P6-Relaxed is thus transformed to the following LP.

$$\mathbf{P7}: \max_{\{\mathbf{x}\}} \sum_{k=1}^{K} \sum_{j=0}^{J} x_{k,j} R_{k,j} \tag{4.40}$$

$$\text{subject to: } (4.11), (4.12), \text{and } (4.39)$$

$$x_{k,j} \in [0,1], \ k = 1, 2, \ldots, K, \ j = 0, 1, \ldots, J.$$

Since **P7** is an LP, the cutting plane method [57] can be applied to obtain its optimal *integer solution*, and such solution is also optimal to **P6** for a given $\mathbf{Q}$.

As users may dynamically join or leave the network, the traffic load of each BS varies over time, the approximation of $\mathbf{Q}$ might be inaccurate. However, a key observation is that *load balancing* can be achieved by solving **P7**. When Q_j is larger than its optimal value, $R_{k,j}$ would be small. Then fewer users would be connected to SBS j after the update with the solution of **P7**, resulting in a decreased Q_j. Thus, the value of Q_j is expected to stay close to its optimal value, and the solution is expected to be near-optimal.

In case the user distribution drastically changes and handover frequently happen (e.g., during rush hours), which can be detected by each BS when measuring the CSI of nearby users, $\mathbf{Q}$ should be updated by solving **P6-Relaxed** with searching over all $\mathbf{Q}$. Due to its high complexity, such update is carried out at a timescale much larger than T.

4.3.3 Iterative Scheme with Near-Optimal Solution

In this section, we propose an iterative approach to obtain the near-optimal solution of the original problem by solving the *WB and MUE resource allocation and pilot length optimization* problem and the *user association* problem iteratively until convergence. The iterative scheme is a three-stage process to guarantee that all constraints are satisfied as well as minimizing the gap of the two problems. The proposed three-stage process is based on the following facts.

Lemma 25 *Under optimal user association solutions, given fixed values of Q_j of other BS's, the sum rate of all users served by SBS j decreases as Q_j increases.*

Proof: According to (4.9), $\sum_{k=1}^{K} R_{k,j}$ is proportional to $\frac{\sum_{k=1}^{K} x_{k,j} \log\left(1+\gamma_{k,j}\right)}{\sum_{k=1}^{K} x_{k,j}}$, which can be interpreted as the average spectral efficiency of users served by SBS j.

Consider an optimal user association with a given feasible set of $\mathbf{Q}$. To maximize the sum rate, the users served by SBS j must be the first Q_j users with the highest spectral efficiencies, i.e., the highest SINRs. Thus, when the values of Q_j for other BS's are fixed, the average spectral efficiency of users served by SBS j decreases as Q_j increases. ■

Property 3 *In* **most cases***, the users served by SBS j are the first Q_j users with highest SINRs, and the sum rate of all users served by SBS j decreases as Q_j increases.*

Compared to Lemma 25, we remove the assumption that the values of Q_j for other BS's are fixed. The only exception of Property 3 happens when a user k' originally served by a neighboring SBS j' is handed over to SBS j due to an increase of $Q_{j'}$, while the SINR of this user is higher than at least one of the users currently served by SBS j. Suppose user k has a lower SINR than user k' when served by SBS j. Then both users are likely to be cell-edge users, and the coverage areas of SBS j and SBS j' are likely to overlap. Hence, the exception case happens when both Q'_j and Q_j increase and a cell-edge user is handed over to SBS j. As a result, when the SBS's are not densely deployed, the exception case would not happen.

Stage I In the first stage, we aim to guarantee that constraints (4.12), $\sum_{k=1}^{K} x_{k,j} \leq S_j$, are always satisfied for all SBS's. We begin with solving the initial MUE and WB resource allocation and pilot length optimization problem without considering the constraint on WB data rate, $\sum_{n=1}^{\alpha N} b_{j,n} = \left\lceil \frac{\sum_{k=1}^{K} x_{k,j} R_{k,j}}{\log(1+\gamma_j)} \right\rceil$. This corresponds to the case of setting the initial values of the RHS of (4.20) to be F_j. Let **P8** be the LP generated by removing constraints (4.12) from **P7**, **P8** can be solved by the same approach as **P7**. Then, we find the optimal user association under WB constraints by solving **P8**. With such initial solution, C_j may be low for SBS j, $\sum_{k=1}^{K} R_{k,j}$ is bounded by a low value. As in Property 3, a large number of users are expected to be assigned to SBS j to achieve a low value of $\sum_{k=1}^{K} R_{k,j}$, which may violate constraint (4.12) and be infeasible to **P7**. Thus, we first solve **P8** to find the set of SBS's that violates the WB constraint, and then enforce additional constraints to **P8** to guarantee feasibility.

With the solution of **P8** , if constraint (4.12) of SBS j is not satisfied, **P8** is updated by adding constraint $\sum_{k=1}^{K} x_{k,j} = S_j$. Then, we update $R_{k,j}$ by keeping the first S_j highest SINR users to be served by SBS j. After that, we update constraint (4.20) for SBS j with the updated $x_{k,j}$ and $R_{k,j}$. This way, both constraints for SBS j are satisfied; the WB resource allocation and user association for SBS j become feasible. Based on Property 3, by keeping the first S_j highest SINR users, the value of $\sum_{k=1}^{K} R_{k,j}$ is expected to be the largest under a feasible and optimal solution of **P7**. This results in the smallest change on the RHS of constraint (4.20) for SBS j. Thus, the change of the polyhedron defined by $\mathbf{Z}$ is minimized, resulting in a smallest reduction of the objective function. Then, we solve the MUE and WB resource allocation and pilot length optimization problem with the updated constraint (4.20) for SBS j. After that, we use the solution to solve **P8** in the next iteration. Such process is repeated until all constraints (4.12) are satisfied for all SBS's. After the process is converged, we enter the second stage.

Stage II In the second stage, we aim to minimize the performance gap between the two problems, so that $C_j - \sum_{k=1}^{K} x_{k,j} R_{k,j}$ is minimized. The motivation of minimizing such gap is because allocating more channels to WBs leads to increased value of τ and decreased data rates of all users, it is desirable that the data rates provided by WBs are sufficiently utilized by each SBS. To minimize the gap at each SBS, we find the SBS's with $\sum_{n=1}^{\alpha N} b_{j,n} > \left\lceil \frac{\sum_{k=1}^{K} x_{k,j} R_{k,j}}{\log(1+\gamma_j)} \right\rceil$, and update these constraints as

$$\sum_{n=1}^{\alpha N} b_{j,n} = \left\lceil \frac{\sum_{k=1}^{K} x_{k,j} R_{k,j}}{\log\left(1+\gamma_j\right)} \right\rceil . \tag{4.41}$$

Then, we obtain the optimal $\{\mathbf{a}, \mathbf{b}, \tau\}$ with the updated constraints as in Section 4.3.1. With $\{\mathbf{a}, \mathbf{b}, \tau\}$, we solve **P7** to obtain the optimal $\mathbf{x}$. Such process is repeated until $\sum_{n=1}^{\alpha N} b_{j,n} > \left\lceil \frac{\sum_{k=1}^{K} x_{k,j} R_{k,j}}{\log(1+\gamma_j)} \right\rceil$ does not hold for any SBS.

Stage III In the third stage, we aim to guarantee that the WB constraints of all SBS's are satisfied after the updates in the second stage. With the update in the second stage, the values of $\sum_{n=1}^{\alpha N} b_{j,n}$ are reduced, which may cause an increased ratio of $\sum_{n=1}^{\alpha N} a_{k,n} / \sum_{n=1}^{\alpha N} b_{j,n}$ for

Algorithm 5: Iterative Scheme to Obtain a Near-Optimal Solution to Problem **P1**

Initialize Set the RHS of (4.20) as F_j ;
do
  Obtain $\{\mathbf{a}, \mathbf{b}, \tau\}$ with Algorithm 4;
  Solve **P8** to obtain $\mathbf{x}$;
  for $j = 1 : J$ **do**
    if $(\sum_{k=1}^{K} x_{k,j} > S_j)$ **then**
      Set $\sum_{k=1}^{K} x_{k,j} = S_j$;
      Update (4.20) for SBS j ;
      Add constraint $\sum_{k=1}^{K} x_{k,j} = S_j$ to **P8** ;
    end
  end
  Solve **P8** to obtain updated $\mathbf{x}$;
  Obtain $\{\mathbf{a}, \mathbf{b}, \tau\}$ with updated $\mathbf{x}$ using Algorithm 4 ;
while $(\sum_{k=1}^{K} x_{k,j} \leq S_j$ *does not hold for all* j *)*;
do
  for $j = 1 : J$ **do**
    if $(\sum_{n=1}^{\alpha N} b_{j,n} > \left\lceil \frac{\sum_{k=1}^{K} x_{k,j} R_{k,j}}{\log(1+\gamma_j)} \right\rceil)$ **then**
      Update $\sum_{n=1}^{\alpha N} b_{j,n}$ with (4.20) ;
    end
  end
  Update $\{\mathbf{a}, \mathbf{b}, \tau\}$ with Algorithm 4 ;
  Update $\mathbf{x}$ by solving **P8** ;
while $(\sum_{n=1}^{\alpha N} b_{j,n} > \left\lceil \frac{\sum_{k=1}^{K} x_{k,j} R_{k,j}}{\log(1+\gamma_j)} \right\rceil$ *holds for any* j *)*;
do
  for $j = 1 : J$ **do**
    if $(\sum_{n=1}^{\alpha N} b_{j,n} \leq \left\lfloor \frac{\sum_{k=1}^{K} x_{k,j} R_{k,j}}{\log(1+\gamma_j)} \right\rfloor)$ **then**
      Update $\sum_{n=1}^{\alpha N} b_{j,n}$ with (4.20) ;
    end
  end
  Update $\{\mathbf{a}, \mathbf{b}, \tau\}$ with Algorithm 4 ;
  Update $\mathbf{x}$ by solving **P8** ;
while $(\sum_{n=1}^{\alpha N} b_{j,n} \leq \left\lfloor \frac{\sum_{k=1}^{K} x_{k,j} R_{k,j}}{\log(1+\gamma_j)} \right\rfloor$ *holds for any* j *)*;

some users. Hence, under the optimal solution of **P8**, these users may switch to the MBS. According to Property 3, the sum rate of SBS's that served these users in the previous iteration are expected to increase, resulting violation of the WB constraints. To deal with this situation, we can adjust and update the values of $\sum_{n=1}^{\alpha N} b_{j,n}$ with (4.20), and we repeat this process until the WB constraints of all SBS's are satisfied.

The procedure of the proposed iterative scheme is summarized in Algorithm 5.

4.3.4 Remarks on Practical Concerns

Quasi-Static Channel Between MBS and SBS's

Due to the fixed locations of SBS's, the channels between SBS's and MBS are quasi-static [52]. As a result, the CSI of WBs can be updated less frequently compared to that of MUEs. This property can be employed to enhance the system performance. In most time periods, the SBS's can use some channels for WB transmission without sending pilots on these channels, resulting in reduced pilot length. Thus, we can assign the WBs to use all the αN channels to increase the data rate. In such scenario, the problem formulation can be derived from Problem **P1** with modifications on the constraints.

Due to the different frequencies of CSI update for MUE and WB, there are two cases at different time periods.

- **First case**: Both WBs and MUEs need to send pilots. When the CSI of WB needs to be updated, all WBs are allocated with one pilot sequence on each channel so that the CSI of WBs on all channels can be obtained. This corresponds to set $b_{j,n} = 1$ for $j = 1, ..., J$, $n = 1, ..., \alpha N$. In addition, the constraint $\sum_{n=1}^{\alpha N} b_{j,n} \leq F_j$ can be removed from Problem **P1** since $\mathbf{b}$ is given.

- **Second case**: Only the MUEs need to send pilots. In these periods, the MBS uses the CSI obtained in the first case until the next update of CSI of WB. Then, the constraint $\sum_{k=1}^{K} a_{k,n} + \sum_{j=1}^{J} b_{j,n} \leq \tau N_{\text{sm}}$ in Problem **P1** should be modified to $\sum_{k=1}^{K} a_{k,n} \leq \tau N_{\text{sm}}$. Since the WBs are allocated with all channels, we have $b_{j,n} = 1$ for $j = 1, ..., J$, $n = 1, ..., \alpha N$. Same as the first case, the constraint $\sum_{n=1}^{\alpha N} b_{j,n} \leq F_j$ is also removed.

With given τ, it can be easily verified that the constraint matrix of the linear programming for solving $\{a_{k,j}\}$ is unimodular for both cases. Thus, we can obtain the optimal $\{\mathbf{a}, \tau\}$ using the same approach as in Algorithm 4. Then, we apply Algorithm 5 to obtain the solutions for both cases.

A Combination of Wired and Wireless Backhaul

In case of dense SBS deployment with heavy traffic load, a long pilot length (i.e., large τ) is required, resulting in degraded system performance. To mitigate such bottleneck as well as preserving the benefits of wireless backhaul, a combination of wired and wireless backhaul is desirable. With proper configuration, a good tradeoff between performance and cost can be achieved.

With a combination of wired and wireless backhaul, the problem formulation needs to be adjusted accordingly. We assume that the data rate of wired backhaul is sufficiently high so that the constraint $\sum_{k=1}^{K} x_{k,j} R_{k,j} \leq C_j$ can always be satisfied. Let Ω be the set of SBS's that uses wireless backhaul, then the constraints $\sum_{k=1}^{K} x_{k,j} R_{k,j} \leq C_j$ and $\sum_{n=1}^{\alpha N} b_{j,n} \leq F_j$ only apply to $j \in \Omega$. The constraints (4.20) and (4.39), which are derived from $\sum_{k=1}^{K} x_{k,j} R_{k,j} \leq C_j$, are also applied to $j \in \Omega$ only. For the constraint $\sum_{k=1}^{K} a_{k,n} + \sum_{j=1}^{J} b_{j,n} \leq \tau N_{\mathrm{sm}}$, we replace the term $\sum_{j=1}^{J} b_{j,n}$ to $\sum_{j \in \Omega} b_{j,n}$. The solution under such new scenario can be obtained with the same approach in Algorithm 5 with the updated constraints (4.20) and (4.39). Specifically, since the SBS's with wired backhaul have no impact on the pilot optimization, we still maximize the sum rate of wireless backhaul and MUE by solving Problem **P2** with Algorithm 4. For user association, the constraint $\sum_{k=1}^{K} x_{k,j} R_{k,j} \leq C_j$ does not apply to the SBS's with wired backhaul, and the problem can be solved with the same approach presented before.

4.4 Distributed Solution Scheme

In the centralized scheme, global network information is required for centralized control, which usually leads to better performance. However, acquiring the global information may incur considerable overhead, which may be infeasible in a large scale network. In this section, we propose a distributed scheme by formulating a noncooperative repeated game among all users. In the repeated game, each user distributively makes its own decision. We demonstrate that the game will converge to an NE.

Algorithm 6: Distributed User Association Strategy for BS j

```
while (convergence not achieved) do
    if (BS j holds more than S_j proposals) then
        Put the top S_j users with the highest SINRs in the waiting list and reject the other users ;
    else
        Put all users in the waiting list ;
    end
end
```

4.4.1 Distributed User Association

We formulate a repeated game among all users, the strategy of each user is to decide its serving BS. Due to the tradeoff in MUE and WB resource allocation, we set a price for using one channel such that the number of channels used by MUEs and WBs can be controlled at proper values. The utility of user k is defined as

$$\begin{cases} \mathcal{U}_{k,0} = \omega_k \log\left(R_{k,0}\right) - p \cdot \sum_{n=1}^{\alpha N} a_{k,n} \\ \mathcal{U}_{k,j} = \omega_k \log\left(R_{k,j}\right) - p \cdot \frac{\sum_{n=1}^{\alpha N} b_{j,n}}{\sum_{k=1}^{K} x_{k,j}},\ j = 0, \ldots, J. \end{cases} \tag{4.42}$$

where ω_k is the evaluation of user k for data rate and p is the price of using one channel. When user k is served by an SBS, the cost of channels for the WB is shared by all users that are served by the SBS. In (4.42), $\sum_{n=1}^{\alpha N} a_{k,n}$ is set by each user to be a fixed value that maximizes its utility, given as $\sum_{n=1}^{\alpha N} a_{k,n} = \arg\max_{\left\{\sum_{n=1}^{\alpha N} a_{k,n}\right\}} \{\mathcal{U}_{k,0}\} = \omega_k / p$. For $\sum_{n=1}^{\alpha N} b_{j,n}$, it is a variable given by (4.41), which is affected by other users' decisions. The strategy of each user is

$$x_{k,j^*} = 1, \quad j^* = \arg\max_{j} \{\mathcal{U}_{k,j}\}. \tag{4.43}$$

To maximize the sum rate under constraint $\sum_{k=1}^{K} x_{k,j} = S_j$, it is reasonable to assume that each BS serves the top S_j users with highest SINRs. The user association strategy of BS's is summarized in Algorithm 6.

Each user has a *preference list* for all BS's, the order of the list is determined by the order of $\mathcal{U}_{k,j}$, e.g., the BS with the largest $\mathcal{U}_{k,j}$ is the first in the preference list of user k. Since Q_j is unknown before the repeated game, the initial preference list of each user is determined

by values of SINRs when connecting to different BS's. The proposed repeated game has the following two stages.

In the *first* stage, each user proposes to the top BS in its preference list. Then, BS's respond to the proposals according to Algorithm 6.

In the *second* stage, each BS j broadcasts the value of Q_j to all users. Then each user k updates its preference list with $R_{k,j}$. A user proposes to another BS under the following cases.

Case 1: The proposal of the user is rejected.

Case 2: A higher utility can be achieved by switching to another BS j' and one of the two conditions is satisfied: (i) $Q_{j'} < S_{j'}$, (ii) $Q_{j'} = S_{j'}$, and there is a user k' currently in the waiting list of BS j' such that $R_{k,j'} > R_{k',j'}$.

If user k is rejected by BS j, it marks BS j as *unavailable* in its preference list. Then, users in these two cases propose to the top BS among remaining available BS's. Once receiving the proposals, each BS compares the new proposals with those in its waiting list, and makes decisions according to Algorithm 6. If a user switches from BS j to BS j' as described in Case 1, the users that once marked BS j as *unavailable* change the status of BS j to *available*. Given the BS decisions, each user then updates its preference list and makes another round of proposal if one of the two cases is satisfied. The repeated game is continued until convergence of user association is achieved.

After convergence, the MBS replaces constraint (4.14) with $\sum_{n=1}^{\alpha N} a_{k,n} = \omega_k/p$ and update constraint (4.15) with (4.20). It then determines $\{\mathbf{a}, \mathbf{b}, \tau\}$ as in Section 4.3.1.

4.4.2 Convergence Analysis

The convergence performance of the repeated game is given in Theorem 6, which shows that an NE can be achieved.

Theorem 6 *The repeated game converges to a Nash equilibrium that is optimal for each user.*

Proof: Suppose the game does not converge. Then, there must be a user k that is currently served by BS j who wishes to propose to another BS j'. Obviously, Case 1 does not hold since user k is served by BS j. Then, Case 2 holds, there is another BS j' such that $\mathcal{U}_{k,j'} > \mathcal{U}_{k,j}$

and BS j' is marked as *available* by user k. If condition (i) is satisfied, $Q_{j'} < S_{j'}$, then user k would have already switched to BS j', which contradicts to the fact that it is served by BS j. If condition (ii) is satisfied, $Q_{j'} = S_{j'}$, then there must another user k' that is served by BS j' such that $R_{k,j'} > R_{k',j'}$, i.e., BS j' prefers user k over user k'. Since user k' is in the waiting list of BS j' while user k is not, it must be the case that user k has never proposed to BS j' before. However, since $\mathcal{U}_{k,j'} > \mathcal{U}_{k,j}$, user k must have proposed to BS j' before BS j, which is also a contradiction. Thus, the repeated game converges.

From the above analysis, we can see that the utility of each user cannot be further improved given the strategies of other users. Thus, the strategy of each user is the *best response* to the strategies of other users when the repeated game converges. We conclude that the repeated game converges to an NE. ■

The order of users that start the proposed process affects the system performance, as different NEs would be achieved. Such randomness results in performance loss of distributed scheme compared to the centralized one.

4.5 Simulation Study

We validate the proposed centralized and distributed schemes with MATLAB simulations. The scenario is based on a cellular system with hexagonal macrocells, and we consider the sum rate of all users in a tagged macrocell area. The MBS is located at the center, the SBS's and users are randomly distributed in the macrocell area. The radius of a macrocell is 500 m. The slow fading factor, $\beta_{k,0}$, is based on the ITU path loss model [59] and a lognormal shadowing with standard deviation of 10 dB. The coherence bandwidth is 150 kHz. We use the parameters of downlink LTE symbol for each OFDM symbol. The spacing between subcarriers is 15 kHz, then $N_{\mathrm{sm}} = 10$; the useful symbol duration $T_u = 1/\Delta_f = 66.7$ ms; and $T_s = T_p = 72$ ms. The coherence time is $T_c = 720$ ms, so each frame has 10 OFDM symbols, and we set $\tau_{\max} = 5$. The total bandwidth is 4 MHz, so the total number of channels is 40. We assume $\alpha = \frac{1}{2}$; then 20 channels are allocated to MUEs and WBs and the other 20 channels are allocated to SUEs. The powers of SBS's are set according to the iterative water-filling scheme [43], with an upper

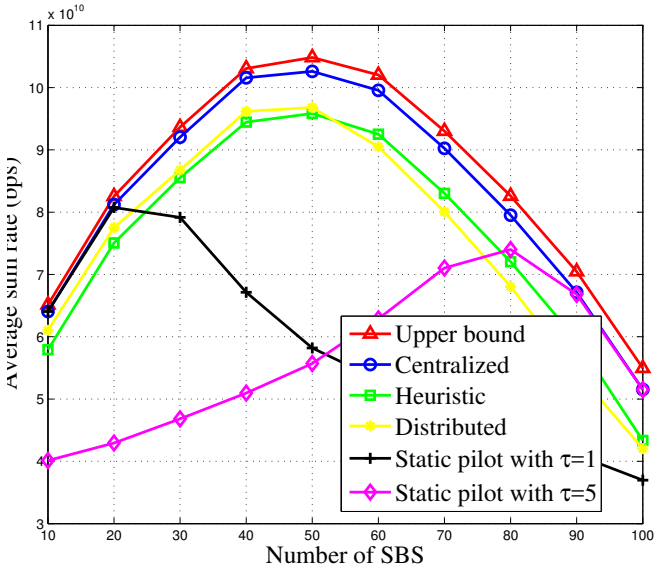

Figure 4.2: Average sum rates of different schemes versus the number of SBS (200 users).

bound of 30 dBm. The upper bounds of $\sum_{k=1}^{K} x_{k,j}$ are set to be $S_j = 20$ for SBS's and $S_0 = 50$ for MBS, respectively.

We compare the proposed schemes with a heuristic scheme, termed *Heuristic*, for user association. Heuristic is based on Property 3 and is derived by making a modification on the centralized scheme. Specifically, instead of solving **P8** at each iteration of the centralized scheme, the set of users served by each SBS is determined with a greedy approach. In each round, we select the user with highest SINR to be served by SBS j and update the value of $\sum_{k=1}^{K} R_{k,j}$. We continue such process until $\sum_{k=1}^{K} R_{k,j} \leq C_j$ is satisfied. We also consider the case based on [87], in which pilot length is not considered for optimization and τ is set as a fixed value (termed *Static pilot*). For Static pilot, the solution of $\{\mathbf{a}, \mathbf{b}, \tau\}$ is based on the solution procedure in Section 4.3.1. For Heuristic, we apply the same procedure of the proposed centralized scheme except the user association strategy. Since the performance of the distributed scheme depends on the value of p, we set p to the value that achieves the maximal sum rate. We also consider the value of the objective function of problem **P2** under optimal solution as an upper bound for comparison.

The sum rate performances of different schemes are presented in Figs. 4.2 and 4.3. In Fig. 4.2, it can be seen that the performances of all schemes first increase and then decrease as the number of SBS's grows. This is because a larger τ is required as the number of SBS's

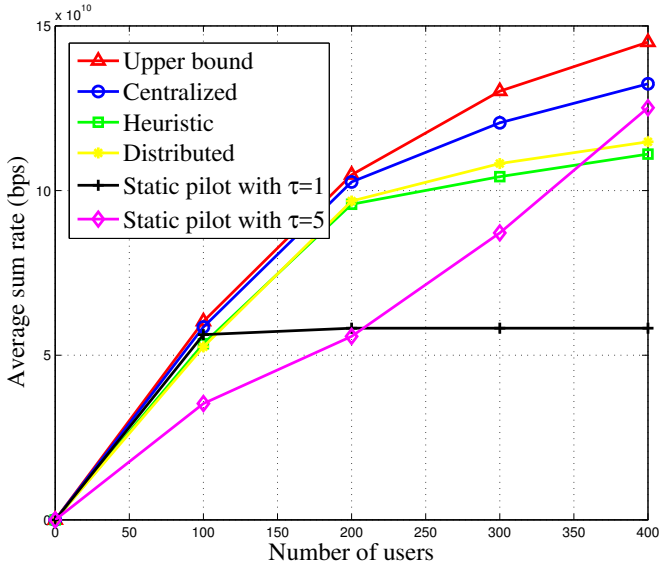

Figure 4.3: Average sum rates of different schemes versus the number of users (20 SBS's).

increases, and the interference between neighboring small cells degrades the average SINRs of SUEs. Both the centralized and distributed schemes outperform Static pilot, demonstrating that a performance gain can be achieved with dynamically adjusted τ. The performance of the centralized scheme is close to its upper bound, since we iteratively minimize the performance gap of two problems in the second stage of the iterative scheme given in Algorithm 5.

It is also observed that the performance of Heuristic is close to that of the centralized scheme when the number of SBS's is small, due to the fact that Property 3 is more reliable when SBS's are not close to each other, and a user would not have close rates by connecting to different SBS's. The distributed scheme also achieves a satisfactory performance since users are charged for using channels, resulting in efficient resource utilization. For Static pilot, the case of $\tau = 1$ achieves better performance than the case of $\tau = 5$ when the number of SBS's is small, a small τ can accommodate the requirements of all WBs. However, when the number of SBS's is large, a larger τ provides better performance since the increased demand for WB data rates can be satisfied.

Fig. 4.3 shows the performances under different numbers of users, where similar trends can be observed. When the number of users increases, the sum rate of users with $\tau = 1$ remains constant. This is because the resources for MUEs and WBs are quite limited. As a result, a considerable proportion of users cannot be served by any BS.

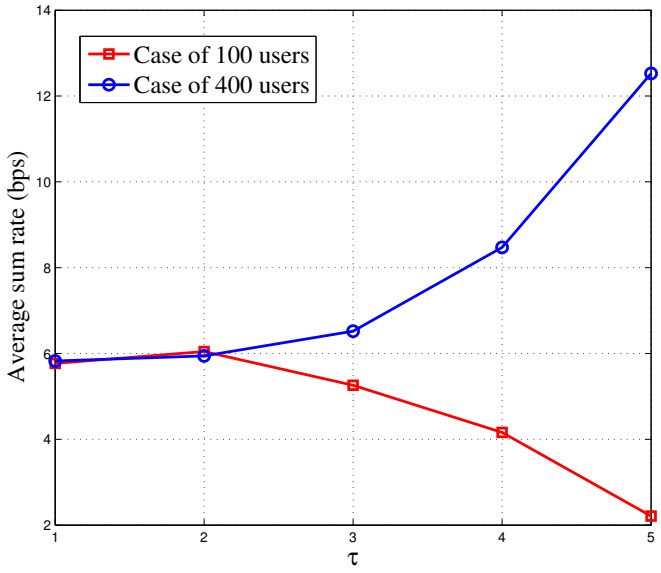

Figure 4.4: Average sum rates versus the value of τ under 2 different numbers of users (20 SBS's).

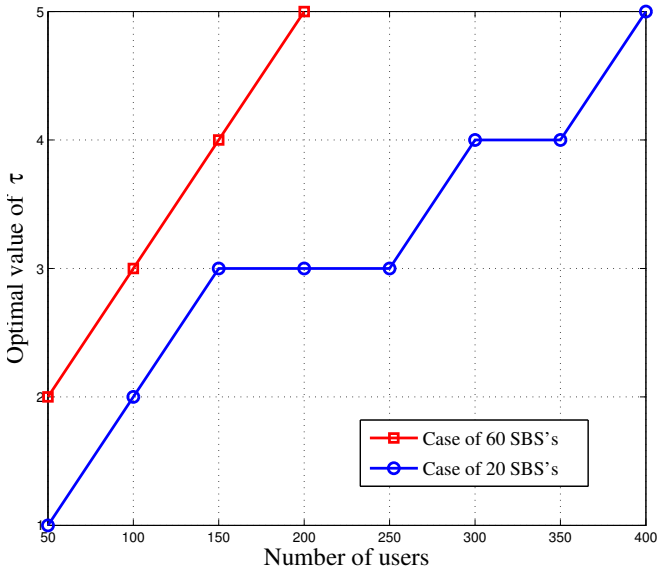

Figure 4.5: Optimal value of τ under different numbers of SBS's (200 users).

In Fig. 4.4, the performance of Static pilot with different τ values is evaluated. When τ is large, the performance with 100 users is significantly worse than that with 400 users; however, when τ is small, the performance with 400 users becomes worse than that with 100 users. This shows that a small value of τ significantly limit the system performance in case of larger number of users, and τ needs to be dynamically adjusted to prevent considerable performance loss of Static pilot under different traffic patterns. The optimal values of τ under different numbers of SBS's and users are presented in Fig. 4.5. The optimal τ increases with both the

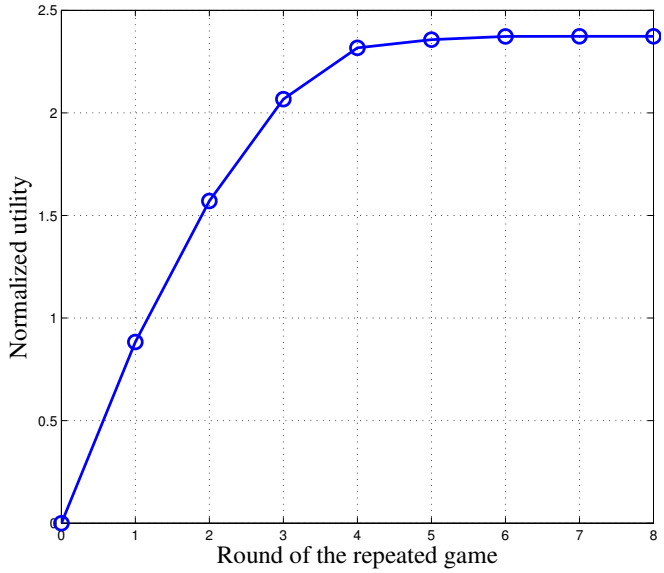

Figure 4.6: Convergence of the repeated bidding game (200 users and 20 SBS's).

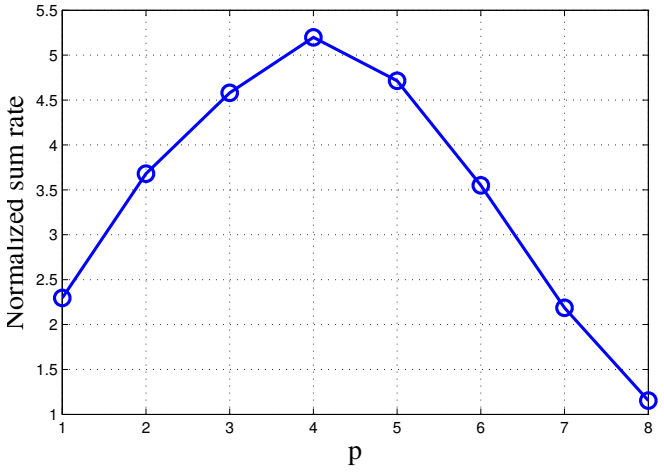

Figure 4.7: Normalized sum rate versus the value of p (200 users and 20 SBS's).

number of SBS's and the number of users, since the more resources are required to satisfy the increasing demand.

An example of the repeated game is given in Fig. 4.6. We can see that the game converges after several rounds and a maximum sum utility is achieved upon convergence.

We also present an example to evaluate the impact of price p on the system performance in Fig. 4.7. By setting p to a proper value, each user makes rational decision on channel usage, the value of τ can be set to a proper value.

4.6 Related Work

The HetNet with WB has been studied in several prior works. Since another type of transmission is added over transmissions between users and BS's, interference management becomes a key issue under certain system assumptions and has been investigated in [97, 98]. In [99], an load-aware design on spatial multiplexing was proposed to improve the energy efficiency of a HetNet with WB. A recent overview on resource management of 5G HetNet with WB was presented in [86].

In this chapter, we integrate massive MIMO into WB and deal with the special challenges with an adaptive frame design. Due to the special architecture of a massive MIMO HetNet with WB, we optimize the network performance with joint frame design, resource allocation, and user association in this chapter.

4.7 Conclusions

In this chapter, we considered the problem of joint frame design, resource allocation, and user association to maximize the sum rate of a massive MIMO HetNet. We formulated a non-linear integer programming problem and proposed a centralized iterative scheme to obtain a near-optimal solution. We also proposed a distributed scheme by formulating a repeated game among all users and prove that the game converges to an NE. Simulation results show that the proposed schemes outperform several benchmark schemes.

Chapter 5

Duplex Mode Selection and Resource Allocation for Full-Duplex Enabled Femtocell Networks

5.1 Introduction

The femtocell technology is initially proposed as an effective solution for enhancing coverage by deploying indoor femtocell base stations (FBS) that are connected through wired links such as cable moderm or Digital Subscriber Line (DSL). The short transmit-receive distance results in high signal to interference plus noise ratio (SINR), and the small coverage area enables dense spatial spectrum reuse, which both contribute to high spectrum utilization. Recently it has been recognized as a key technology by Qualcomm for meeting the 1000x data challenge, i.e., the predicted astounding 1000x increase in mobile data in the near future [1]. However, due to the current spectrum scarcity problem, femtocells are more likely to operate on the same spectrum band with the existing macrocells, resulting in cross-tier interference (between femtocells and macrocells) and inter-femtocell interference (among femtocells). Interference management is critical for the success of this technology.

The cognitive femtocell network (CFN) is proposed as a solution to the interference problem [107,108]. In general, the macrocell users (MU) are regarded as primary users (PU) and the femtocell users (FU) are regarded as secondary users (SU). The FBS's periodically sense the spectrum usage of MUs and allocate the unoccupied channels to FUs. The previous research works aim to improve the performance (such as throughput, capacity, energy efficiency, etc.) of CFN as well as guaranteeing the QoS of both MUs and FUs. In [80, 109], spectrum and power allocations in a CFN are formulated as optimization problems, with the objectives to maximize capacity and energy efficiency, respectively. In [110], a strategic game model was introduced

by setting the payoff of a femtocell as the expected number of resource blocks (RB) without interference. With this mechanism, each femtocell makes rational decisions on the spectrum usage pattern and the interference between femtocells is mitigated. Another game theoretic mode was proposed in [111], where the penalty of a femtocell is determined by excessive usage of RBs and transmission power. The femtocells are thus discouraged to occupy excessive RBs and transmit with high power, resulting in mitigated interference.

A general approach to address the interference problem is to restrict the spectrum and power usage of femtocells. However, when the number of MUs or the number of femtocells in the CFN is large (e.g., in a hotspot), the spectrum allocated to each femtocell could be limited. As the femtocell technology is expected to provide high data rate services to FUs, the limited spectrum resource may be insufficient to guarantee their QoS. To remedy this disadvantage, more efficient spectrum reuse is required. With the recent development of self-interference suppression technology, a wireless transceiver is able to simultaneously transmit and receive signals on the same channel, yielding a full-duplex (FD) transmission pattern [114]. Theoretically, an FD transmission could double the system capacity, making it a promising approach to improve spectrum utilization. In [115], an FD OFDMA based multi-cell network was investigated, in which the FD empowered BS simultaneously serves two cellular users on the same channel. Despite the presence of inter-cell and intra-cell interference, the results show that the capacity can be enhanced by 86% in the uplink and 99% in the downlink.

The successful use of FD in cellular network motivates us to integrate this technology into the CFN. This is a more challenging case due to the more complicate interference scenarios in a CFN. Similar to the cellular network, an FBS in the CFN can simultaneously serve a pair of FUs on the same channel, resulting in the improved spectrum utilization. However, due to the limited processing capability and battery capacity of mobile device, self-interference suppression may not be applicable to femtocell user equipments (FUE). For the two users that use the same channel, the uplink signal of one user causes interference to the downlink of the other user. To control such intra-femtocell interference, it is necessary to carefully schedule the FUs that are paired for FD transmission. When the intra-femtocell interference is strong between FUEs, the half-duplex (HD) would be a better choice. Therefore, the duplex mode

selection strategy and channel allocation of femtocells should be carefully designed to achieve high capacity as well as mitigating intra-femtocell interference.

In this chapter, we consider a CFN integrated with FD functionality. In such a full-duplex cognitive femtocell network (FDCFN), we aim to maximize the sum rate of FUs as well as guaranteeing the QoS of both FUs and MUs in the form of a minimum SINR requirement. The goals are achieved through duplex mode selection, distributed power control, and channel allocation. The main contributions of this chapter are summarized as follows.

- We incorporate the FD and CR technologies into femtocell networks, and develop a holistic formulation of the joint duplex mode selection, power control, and channel allocation problem in an FDCFN.

- We first consider power control over a pair of channels, and propose two optimal power allocation schemes that can be used in sparse and dense femtocell deployment scenarios, respectively.

- For the case of multiple channels, we propose an iterative framework that jointly solve the duplex selection, power control, and channel allocation problems, and obtain the near optimal solution. We also prove the guaranteed convergence of the proposed framework.

- We propose a duplex mode selection strategy for FDCFN. The FUE pairing is formulated as a roommate matching problem, and we develop an effective algorithm to solve the matching problem. The duplex mode selection is based on the pairing result to achieve high capacity gains.

- We employ the SCALE (Successive Convex Approximation for Low-complExity) algorithm to solve the power control in FDCFN with a distributed approach.

- We propose a greedy channel allocation algorithm for the FDCFN based on the pairing result and derive a performance lower bound.

- The proposed schemes are evaluated with simulations and comparison with several benchmark schemes, where superior performance of the proposed schemes is observed.

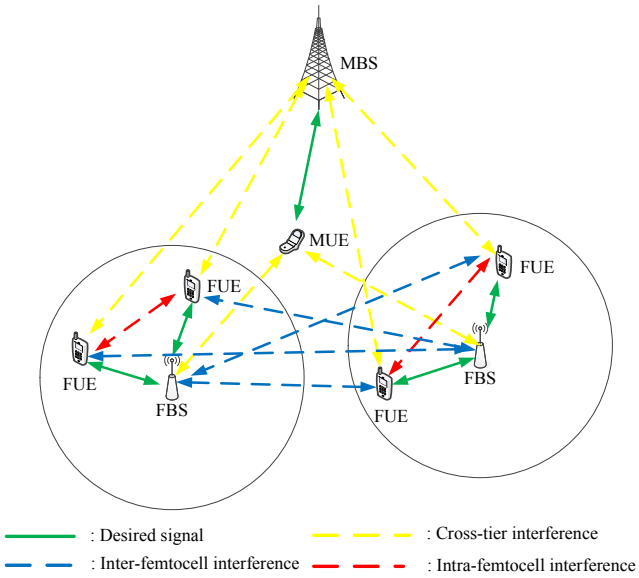

Figure 5.1: The system model for an FD cognitive femtocell network.

The remainder of this chapter is organized as follows. The problem formulation is described in Section 5.2. The power control over a pair of channels is discussed in Section 5.3. The joint duplex mode selection, power control, and channel allocation over multiple channels is investigated in Section 5.4. The performance evaluation is presented in Section 5.5. Section 5.6 discusses related works and Section 5.7 concludes the chapter.

5.2 Problem Formulation

We consider a cognitive femtocell network with one MBS and F femtocells, as shown in Fig. 5.1. The FUs are treated as SUs while the MUs act as PUs. All the femtocells operate on the same spectrum band as the macrocell. Both the macrocell and femtocells are based on OFDMA, where a channel consists of several sub-carriers with bandwidth W. Without loss of generality, we assume that only the FBS's can operate in the FD mode, while the FUEs cannot. The macrocell adopts frequency division duplexing (FDD), i.e., the MBS assigns two channels for an MUE for uplink and downlink transmissions, respectively.

5.2.1 SINR Analysis

Let binary variables $a^u_{f,i}(n), a^d_{f,i}(n) \in \{0,1\}$ be the channel allocation indicators, where $a^u_{f,i}(n) = 1$ ($a^d_{f,i}(n) = 1$) indicates that channel n is allocated to FUE i in femtocell f for

uplink (downlink) transmission, and $a^u_{f,i}(n) = 0$ (or, $a^d_{f,i}(n) = 0$) otherwise. The corresponding transmit powers on the channels are denoted as $p^u_{f,i}(n)$ and $p^d_{f,i}(n)$, respectively.

Let $\gamma_{MUE}(n)$ denote the received SINR of an MUE on channel n, which is given by

$$\gamma_{MUE}(n) = \frac{p_b H_{b,m}(n)}{I_f(n) + N_0}, \tag{5.1}$$

where p_b is the MBS transmit power on the channel, $H_{b,m}(n)$ is the channel gain between the MBS and the MUE, and N_0 is the noise power on the channel. Let π_f be the set of FUEs in femtocell f. Let $H^u_{f,i,m}(n)$ be the channel gain between FUE i in femtocell f and MUE m on channel n, $H^d_{f,i,m}(n)$ be the channel gain between FBS f and MUE m on channel n. Denote $I_f(n)$ as the interference caused by all the femtocell transmissions on channel n, as

$$I_f(n) \;=\; \sum_{f=1}^{F} \sum_{i \in \pi_f} \{a^u_{f,i}(n) p^u_{f,i}(n) H^u_{f,i,m}(n) + a^d_{f,i}(n) p^d_{f,i}(n) H^d_{f,i,m}(n)\}. \tag{5.2}$$

Similarly, let p_m be the transmit power of the MUE. Assuming channel reciprocity, the SINR at the MBS for the MUE on channel n, denoted as $\gamma_{MBS}(n)$, is given by

$$\gamma_{MBS}(n) = \frac{p_m H_{b,m}(n)}{I_f(n) + N_0}. \tag{5.3}$$

As shown in Fig. 5.1, there are three types of interference in an FDCFN, namely the cross-tier interference, inter-femtocell interference, and intra-femtocell interference. The SINR at FUE i in femtocell f on channel n is given by

$$\gamma^{FUE}_{f,i}(n) = \frac{p^d_{f,i}(n) H^d_{f,i}(n)}{I_{of}(n) + I_{ff}(n) + I_{pf}(n) + N_0}, \tag{5.4}$$

where $H^d_{f,i}(n)$ is channel gain between FUE i and FBS f on channel n. $I_{of}(n)$ is the interference caused by FUEs and FBS's in other femtocells operating on channel n, which can be

derived as

$$I_{of}(n) = \sum_{k=1,k\neq f}^{F} \sum_{l\in\pi_k} \{a^u_{k,l}(n)p^u_{k,l}(n)H^u_{f,i,k,l}(n) + a^d_{k,l}(n)p^d_{k,l}(n)H^d_{f,i,k}(n)\}, \tag{5.5}$$

where $H^u_{f,i,k,l}(n)$ is the channel gain between FUE l in femtocell k and FUE i in femtocell f on channel n, and $H^d_{f,i,k}(n)$ is the channel gain between FBS k and FUE i in femtocell f on channel n. Denote $I_{ff}(n)$ as the intra-femtocell interference caused by FUE j to FUE i. Assuming that FUE j and FUE i are paired to operate in the FD mode, $I_{ff}(n)$ can be written as

$$I_{ff}(n) = a^u_{f,j}(n)p^u_{f,j}(n)H_{f,i,j}(n), \tag{5.6}$$

where $H_{f,i,j}$ is the channel gain between FUE i and FUE j on channel n. $I_{pf}(n)$ is the cross-tier interference caused by the MUE or MBS using channel n, which is

$$I_{pf}(n) = p_m H^u_{f,i,m}(n) \text{ or } I_{pf}(n) = p_b H^u_{f,i,b}(n), \tag{5.7}$$

where $H^u_{f,i,m}$ and $H^u_{f,i,b}$ are the channel gains between FUE i in femtocell f and MUE m, and the MBS on channel n, respectively.

The SINR at FBS f on channel n is given by

$$\gamma^{FBS}_{f,i}(n) = \frac{p^u_{f,i}(n)H^u_{f,i}(n)}{I_{of}(n) + I_{sf}(n) + I_{pf}(n) + N_0}. \tag{5.8}$$

$I_{sf}(n)$ is the residual self-interference on channel n at the FBS. The self-interference suppression coefficient $0 < \kappa < 1$ is defined as the ratio of residual self-interference power to the original self-interference power. $I_{sf}(n)$ is given as

$$I_{sf}(n) = \kappa p^d_{f,j}(n). \tag{5.9}$$

5.2.2 Sum Rate Maximization

Based on the SINR analysis, the sums of the achievable rates for the uplink and downlink are given by

$$\mathcal{C}_{FUE} = \sum_{f=1}^{F}\sum_{n=1}^{N} a_{f,i}^{d}(n) W \log_2(1+\gamma_{f,i}^{FUE}(n)), \tag{5.10}$$

$$\mathcal{C}_{FBS} = \sum_{f=1}^{F}\sum_{n=1}^{N} a_{f,i}^{u}(n) W \log_2(1+\gamma_{f,i}^{FBS}(n)), \tag{5.11}$$

respectively. We formulate the sum rate maximization problem for the FD cognitive femtocell network as follows.

$$\underset{\{a_{f,i}^{u}(n), a_{f,i}^{d}(n), p_{f,i}^{u}(n), p_{f,i}^{d}(n)\}}{\arg\max} \{\mathcal{C}_{FBS} + \mathcal{C}_{FUE}\} \tag{5.12}$$

subject to:

$$\gamma_{MUE}(n) \geq \Gamma_1, \gamma_{MBS}(n) \geq \Gamma_1, \ \forall\, n, \tag{5.13}$$

$$\gamma_{f,i}^{FUE}(n) \geq \Gamma_2, \gamma_{f,i}^{FBS}(n) \geq \Gamma_2, \ \forall\, n, f, i \in \pi_f, \tag{5.14}$$

$$a_{f,i}^{u}(n) + a_{f,j}^{u}(n) \leq 1, \ \forall\, n, f, i, j \in \pi_f, \tag{5.15}$$

$$a_{f,i}^{d}(n) + a_{f,j}^{d}(n) \leq 1, \ \forall\, n, f, i, j \in \pi_f, \tag{5.16}$$

$$a_{f,i}^{u}(n) + a_{f,j}^{d}(n) \leq 2, \ \forall\, n, f, \forall\, i, j \in \pi_f. \tag{5.17}$$

$$\sum_{n=1}^{N} p_{f,i}^{u}(n) \leq P_{m1}, \ \forall f, \forall i \in \pi_f, \tag{5.18}$$

$$\sum_{n=1}^{N} p_{f,i}^{d}(n) \leq P_{m2}, \ \forall f, \forall i \in \pi_f. \tag{5.19}$$

In (5.13) and (5.14), Γ_1 and Γ_2 are the minimal SINRs to satisfy the QoS requirements of the macrocell and femtocell operations, respectively. Inequalities (5.15) and (5.16) are due to the fact that a channel cannot be shared by two FUEs for uplink or downlink transmissions. Inequality (5.17) is because in the best case, a channel can be shared by two FUs: one FU uses the channel for uplink transmission and the other FU uses the channel for downlink transmission.

In power constraints (5.18) and (5.19), P_{m1} and P_{m2} are the maximal powers of an FUE and an FBS, respectively.

5.3 Optimal Power Control Scheme over a Pair of Channels

In this section, we consider the case of a pair of channels, which are shared by a pair of FUEs. Problem (5.12) is then reduced to power control over these two channels. In the following, we discuss two cases and develop efficient methods to this power control problem.

5.3.1 Case of Sparse Femtocell Deployment

We first consider the case of sparse femtocell deployment, where femtocells are apart from each each with enough distance so that inter-femtocell interference can be neglected. Since the inter-femtocell interference is neglected, the powers of different FBS's and FUEs do not impact with each other, so power control can be independently performed for each femtocell. Thus, the problem is to find the optimal power control scheme to maximize the sum rate of each femtocell. We assume that with proper channel allocation (which will be discussed in Section 5.4.3), FBS's and FUEs utilize different channels with nearby MUEs or MBS, such that the QoS requirements of MUs are satisfied. We also assume that with proper user pairing method (which will be described in Section 5.4.1), the intra-femtocell interference is effectively controlled. Thus, with the short distance of femtocell transmissions, the FBS and FUE are expected to operate in the high SINR region.

Without loss of generality, we consider two FUs, FU 1 and FU 2. One channel is used by FU 1 for uplink transmissions and by FU 2 for downlink transmissions; the other channel is used by FU 1 for downlink transmissions and by FU 2 for uplink transmissions. Denote p_1^u and p_2^u as the uplink powers, and p_1^d and p_2^d the downlink powers of of the two FUEs. H_{11} and H_{22} are the channel gains between the FBS and the two FUEs, respectively, and H_{12} the channel gain between two FUEs. Let I_{p_0} be the interference power from primary users (i.e. the MBS or MUE) at the FBS, and I_{p_1} and I_{p_2} be the interference power at FUE 1 and FUE 2, respectively.

In the high SINR region, the original problem (5.12) can be approximated as

$$\begin{aligned}\underset{\{p_1^u,p_1^d,p_2^u,p_2^d\}}{\arg\max}\ \Big\{&\log_2\left(\tfrac{p_1^d H_{11}}{I_{p_1}+p_2^u H_{12}+N_0}\right)+\log_2\left(\tfrac{p_2^d H_{22}}{I_{p_2}+p_1^u H_{12}+N_0}\right)\\ &+\log_2\left(\tfrac{p_1^u H_{11}}{I_{p_0}+p_2^d\kappa+N_0}\right)+\log_2\left(\tfrac{p_2^u H_{22}}{I_{p_0}+p_1^d\kappa+N_0}\right)\Big\}.\end{aligned} \tag{5.20}$$

Since only a pair of channels are considered, inter-femtocell and cross-tier interference are neglected, the constraints become

$$p_1^u \le P_{m1}, \quad p_2^u \le P_{m1}, \tag{5.21}$$

$$p_1^d + p_2^d \le P_{m2}. \tag{5.22}$$

Although the objective function (5.20) is non-convex, it can be rewritten as

$$\underset{\{p_1^u,p_1^d,p_2^u,p_2^d\}}{\arg\max}\ \log_2\left(\frac{H_{11}^2 H_{22}^2}{T\left(p_1^u,p_1^d,p_2^u,p_2^d\right)}\right), \tag{5.23}$$

where

$$\begin{aligned}T\left(p_1^u,p_1^d,p_2^u,p_2^d\right) &= \frac{\left(N_0+I_{p_0}\right)^2\left(N_0+I_{p_1}\right)\left(N_0+I_{p_2}\right)}{p_1^u p_1^d p_2^u p_2^d}+\\ &\frac{\left(N_0+I_{p_0}\right)^2\left(N_0+I_{p_1}\right)H_{12}}{p_1^d p_2^u p_2^d}+\frac{\left(N_0+I_{p_0}\right)^2\left(N_0+I_{p_2}\right)H_{12}}{p_1^u p_1^d p_2^d}+\\ &\left(N_0+I_{p_0}\right)\left(N_0+I_{p_1}\right)\left(N_0+I_{p_2}\right)\kappa\left(\frac{1}{p_1^u p_2^u p_2^d}+\frac{1}{p_1^u p_2^u p_1^d}\right)+\\ &\frac{\left(N_0+I_{p_0}\right)^2 H_{12}^2}{p_1^d p_2^d}+\\ &\left(N_0+I_{p_0}\right)\left(N_0+I_{p_1}\right)H_{12}\kappa\left(\frac{1}{p_2^u p_2^d}+\frac{1}{p_2^u p_1^d}\right)+\\ &\frac{\left(N_0+I_{p_1}\right)\left(N_0+I_{p_2}\right)\kappa^2}{p_1^u p_2^u}+\left(N_0+I_{p_0}\right)H_{12}^2\kappa\left(\frac{1}{p_2^d}+\frac{1}{p_1^d}\right)+\\ &\left(N_0+I_{p_0}\right)\left(N_0+I_{p_2}\right)H_{12}\kappa\left(\frac{1}{p_1^u p_2^d}+\frac{1}{p_1^u p_1^d}\right)+\\ &\frac{\left(N_0+I_{p_1}\right)H_{12}\kappa^2}{p_2^u}+\frac{\left(N_0+I_{p_2}\right)H_{12}\kappa^2}{p_1^u}+H_{12}^2\kappa^2.\end{aligned} \tag{5.24}$$

Since the logarithmic function is monotonically increasing, maximizing (5.20) is equivalent to minimizing $T\left(p_1^u, p_1^d, p_2^u, p_2^d\right)$.

Lemma 26 *$T\left(p_1^u, p_1^d, p_2^u, p_2^d\right)$ is a strictly convex function over $p_1^d, p_1^u, p_2^d, p_2^u$.*

Proof: $T\left(p_1^u, p_1^d, p_2^u, p_2^d\right)$ is a weighted sum of functions $f\left(x_1, \cdots, x_n\right) = 1/\prod_{i=1}^{n} x_i, x_1, x_2, \ldots, x_n > 0$. Let $\mathbf{X}_{n\times n}$ be the Hessian matrix of $f\left(x_1, \cdots, x_n\right)$, given as:

$$\mathbf{X}_{n\times n} = \frac{1}{x_1 \cdots x_n} \begin{pmatrix} \frac{1}{x_1^2} & \cdots & \frac{1}{x_1 x_n} \\ \vdots & \ddots & \vdots \\ \frac{1}{x_1 x_n} & \cdots & \frac{1}{x_n^2} \end{pmatrix}. \tag{5.25}$$

Let $\mathbf{y} = [y_1, \cdots y_n]^T, \mathbf{y} \neq \mathbf{0}$ be an arbitrary non-zero vector, then $\mathbf{y}^T\mathbf{X}\mathbf{y}$ can be calculated by:

$$\mathbf{y}^T\mathbf{X}\mathbf{y} = \frac{\sum_{i=1}^{n}\left(\frac{y_i}{x_i}\right)^2 + \left(\sum_{i=1}^{n}\frac{y_i}{x_i}\right)^2}{\prod_{i=1}^{n} x_i} > 0, i \in \{1, \ldots, n\}. \tag{5.26}$$

Therefore, we conclude that the Hessian matrix is positive definite, and $f\left(x_1, \cdots, x_n\right)$ is a convex function. Since $T\left(p_1^u, p_1^d, p_2^u, p_2^d\right)$ is a weighted sum of convex functions, it is also a convex function. ■

Obviously, both constraints (5.21) and (5.22) are linear. Thus, minimizing (5.24) with constraints (5.21) and (5.22) is a convex optimization problem. Let $V\left(p_1^u, p_1^d, p_2^u, p_2^d, \nu_1, \nu_2, \nu_3\right)$ be the Lagrangian function with Lagrange multipliers ν_1, ν_2, and ν_3 corresponding to the three constraints, respectively. Applying the KKT conditions, we have

$$\begin{cases} \frac{\partial V}{\partial p_1^d} = \frac{\partial V}{\partial p_1^u} = \frac{\partial V}{\partial p_2^d} = \frac{\partial V}{\partial p_2^u} = 0 \\ \nu_1\left(P_{m1} - p_1^u\right) = \nu_2\left(P_{m1} - p_2^u\right) = 0 \\ \nu_3\left(P_{m2} - p_1^d - p_2^d\right) = 0. \end{cases} \tag{5.27}$$

We derive the optimal solution as

$$\begin{cases} p_1^u = p_2^u = P_{m1}, \\ p_1^d = p_2^d = \frac{P_{m2}}{2}. \end{cases} \tag{5.28}$$

It is counter-intuitive that the optimal powers are independent of the channel gains. This is because the objective is to maximize the sum uplink and downlink rates of two users, and the powers present a symmetric pattern in the objective function.

5.3.2 Case of Dense Femtocell Deployment

With dense deployment pattern, femtocells are close to each other, inter-femtocell interference become a major factor that impacts the system performance. Therefore, the FBS and FUE may not always operate in the high SINR region, and the approximation used in Section 5.3.1 may not be applicable. To this end, we leverage the SCALE (Successive Convex Approximation for Low-complExity) algorithm [120] to transform the power control problem into a convex problem. As in Section 5.3.1, we assume that with proper channel allocation scheme, the cross-tier interference and inter-femtocell interference are controlled within acceptable levels such that the QoS requirements of both MUs and FUs are satisfied.

Let channel n_1 and n_2 be the pair of channels to be considered. Denote Π_{n_1} and Π_{n_2} as the set of FBS-FUE links that operate on channel n_1 and n_2, respectively. Then, the sum rate maximization problem of all femtocells on channel n_1 and n_2 can be formulated as

$$\underset{\{p_{f,i}^u(n), p_{f,i}^d(n)\}}{\arg\max} \sum_f \sum_n \Big\{ \log_2 \left(1 + \gamma_{f,i}^{FUE}(n)\right) + \log_2 \left(1 + \gamma_{f,i}^{FBS}(n)\right) \Big\} \tag{5.29}$$

$$\text{subject to: } n \in \{n_1, n_2\} \tag{5.30}$$

$$\{f, i\} \in \{\Pi_{n_1} \cup \Pi_{n_2}\} \tag{5.31}$$

Power constraints (5.18) and (5.19),

where $\gamma_{f,i}^{FUE}(n)$ and $\gamma_{f,i}^{FBS}(n)$ are given by (5.4) and (5.8), respectively. The sets Π_{n_1} and Π_{n_2} are determined by $a_{f,i}^{u}(n)$ and $a_{f,i}^{d}(n)$. With the assumption that the QoS requirements of MUs and FUs are satisfied through channel allocation, only the power constraints (5.18) and (5.19) are considered in this part.

The objective function of (5.29) can be expressed as the difference of two concave functions, which is a well-known NP-hard problem. To deal with such problems, an effective approximation for the function $\log_2(1+z)$ had been employed to transform the original problem to a convex approximation [120–122]. For $z \geq 0$, a lower bound of the logarithm function is given by

$$\log_2(1+z) \geq \alpha \log_2 z + \beta. \tag{5.32}$$

The coefficients are given by:

$$\begin{cases} \alpha = \frac{z_0}{1+z_0} \\ \beta = \log_2(1+z_0) - \frac{z_0}{1+z_0}\log_2 z_0 \end{cases} \tag{5.33}$$

where $z_0 \geq 0$. It can be easily verified that this lower bound is tight when $z = z_0$. Thus, we can maximize the lower bound of (5.29), then tighten this bound by iteratively updating α and β according to the newly calculated SINR values. The lower bound can be expressed as

$$\begin{aligned} B = \sum_{f}\sum_{n} \{ & \alpha_{f,i}^{FUE}(n) \log_2\left(\gamma_{f,i}^{FUE}(n)\right) + \beta_{f,i}^{FUE}(n) + \\ & \alpha_{f,i}^{FBS}(n) \log_2\left(\gamma_{f,i}^{FBS}(n)\right) + \beta_{f,i}^{FBS}(n) \}. \end{aligned} \tag{5.34}$$

To simplify notation, we use matrices in the following. Denote $\boldsymbol{\alpha}^{[t]}$, $\boldsymbol{\beta}^{[t]}$, and $\boldsymbol{\gamma}^{[t]}$ as the matrices for lower bound coefficients and SINR values at the tth iteration, respectively. Let $\mathbf{p}_u^{[t]}$ and $\mathbf{p}_d^{[t]}$ be the matrices for the uplink and downlink powers, respectively, and $\mathbf{p}^{[t]}$ the matrix for all the powers.

Note that (5.34) is still the difference of concave functions. However, we can apply a logarithmic transform of variables by defining substitution variables $\mathbf{q}^{[t]} = \ln\{\mathbf{p}^{[t]}\}$, to convert

Algorithm 7: SCALE Algorithm

Initialize $t = 1$, $\boldsymbol{\alpha}^{[1]} = 1$, $\boldsymbol{\beta}^{[1]} = 0$;
do
  Solve (5.35) to obtain $\mathbf{q}_u^{[t]}$, $\mathbf{q}_d^{[t]}$;
  Compute $\mathbf{p}_u^{[t]} = e^{\mathbf{q}_u^{[t]}}$, $\mathbf{p}_d^{[t]} = e^{\mathbf{q}_d^{[t]}}$, $\boldsymbol{\gamma}^{[t]}$;
  Update $\boldsymbol{\alpha}^{[t+1]}$, $\boldsymbol{\beta}^{[t+1]}$ using $\boldsymbol{\gamma}^{[t]}$ as in (5.33) ;
  $t \leftarrow t + 1$;
while $|\gamma^{[t+1]} - \gamma^{[t]}| \geq \epsilon$;

it into a convex problem. The maximization of the lower bound for the tth iteration can be expressed as

$$\arg\max_{\{\mathbf{q}^{[t]}\}} B\left(\mathbf{q}^{[t]}, \boldsymbol{\alpha}^{[t]}, \boldsymbol{\beta}^{[t]}\right) \tag{5.35}$$

$$\text{subject to: } \mathbf{q}_u^{[t]} \leq \ln\{\mathbf{P}_{m1}\} \tag{5.36}$$

$$\sum_{n=n_1,n_2} \exp\left\{\mathbf{q}_d^{[t]}\right\} \leq \mathbf{P}_{m2}. \tag{5.37}$$

Since (5.35) is a convex problem [120], it can be effectively solved in the dual domain. The dual problem is given by

$$\min_{\{\boldsymbol{\lambda},\boldsymbol{\mu}\}} g\left(\boldsymbol{\lambda}, \boldsymbol{\mu}\right), \tag{5.38}$$

where $\boldsymbol{\lambda}$ and $\boldsymbol{\mu}$ are Lagrange multipliers and $g\left(\boldsymbol{\lambda}, \boldsymbol{\mu}\right)$ is the Lagrangian dual function defined as

$$g\left(\boldsymbol{\lambda}, \boldsymbol{\mu}\right) = \max_{\{\mathbf{q}_u, \mathbf{q}_d, \boldsymbol{\lambda}, \boldsymbol{\mu}\}} \Big\{B + \boldsymbol{\lambda}^T \left(\ln\{\mathbf{P}_{m1}\} - \mathbf{q}_u\right) + \boldsymbol{\mu}^T \left(\mathbf{P}_{m2} - \sum_{n\in\{n_1,n_2\}} \exp\left\{\mathbf{q}_d\right\}\right)\Big\}. \tag{5.39}$$

Since problem (5.35) is a special case for the problem that will be discussed in Section V.4-B, we will present the detail procedures to solve the problem is Section 5.4.3 With the solution to problem (5.35) solved, we then apply the SCALE algorithm to obtain the final solution to the power control problem, as presented in Algorithm 7.

Denote T_s as the number of iterations required to solve (5.35) in the dual domain. Given the channel allocation result, the complexity of each iteration of the SCALE algorithm is $O(T_s F N)$ (with $N = 2$ in this section). The convergence property of the SCALE algorithm is given by the following lemma.

Lemma 27 *The objective function of the SCALE algorithm monotonically increases at each iteration, and finally converges.*

Proof: Let $C\left(\mathbf{p}^{[t]}\right)$ be the objective value of (5.29) at the tth iteration. The following inequalities and equalities hold.

$$\begin{aligned} &\cdots C\left(\mathbf{p}^{[t]}\right) \overset{(1)}{=} B\left(\mathbf{q}^{[t]}, \boldsymbol{\alpha}^{[t]}, \boldsymbol{\beta}^{[t]}\right) \overset{(2)}{\leq} B\left(\mathbf{q}^{[t+1]}, \boldsymbol{\alpha}^{[t]}, \boldsymbol{\beta}^{[t]}\right) \\ &\overset{(3)}{\leq} C\left(\mathbf{p}^{[t+1]}\right) \overset{(1)}{=} B\left(\mathbf{q}^{[t+1]}, \boldsymbol{\alpha}^{[t+1]}, \boldsymbol{\beta}^{[t+1]}\right) \cdots. \end{aligned} \tag{5.40}$$

Equality (1) is because the lower bound is tight by setting $\boldsymbol{\alpha}^{[t]}$ and $\boldsymbol{\beta}^{[t]}$ according to the current SINR values. Inequality (2) is due to the fact that problem (5.35) is convex and the updated powers either increase or maintain the objective value. Inequality (3) is derived from (5.32). Since $C\left(\mathbf{p}^{[t]}\right)$ is finite, the SCALE algorithm converges. ■

5.4 Duplex Mode Selection, Power Control and Channel Allocation over Multiple Channels

In this section, we consider joint duplex mode selection, power control and channel allocation over multiple channels. We propose an iterative framework to solve problem (5.12), by decomposing the original problem into three subproblems: duplex mode selection, power control, and channel allocation, and solving each subproblem by fixing the solution of the other two subproblems. The subproblems are iteratively solved until the solution converges. We prove that the proposed iterative framework converges.

5.4.1 Duplex Mode Selection and FUE Pairing based on Stable Roommate Matching

To achieve FD transmissions, an FBS needs to schedule a pair of FUEs to simultaneously operate on the same channel: with one FUE using the channel for uplink transmission and

the other FUE using the channel for downlink transmission [115]. Although the FBS can adopt effective self-interference cancellation, the intra-femtocell interference caused by the uplink of one FUE to the downlink of the other FUE remains a critical problem. Under some circumstances, such intra-femtocell interference can severely degrade the QoS of the FUs. As a result of the low data rates, the FD mode would be inefficient compared to the traditional HD mode.

To fully harvest the potential of FD transmission in the presence of interference between FUEs, a desirable approach is to schedule two users who are relatively far from each other to use the same channel. When multiple FUEs are served by an FBS, it is necessary to design a pairing strategy to find the pair of FUEs for the FD mode. From the perspective of an FUE, there are preferred and undesired FUEs to be paired with, since pairing with different FUEs results in different interference and QoS. This observation motivates us to utilize the model of stable roommate matching to characterize the FUE pairing problem [126].

In the stable roommate matching problem, we consider a group of people who wish to find a satisfactory roommate. Each person has a preference list selected from all other people in the group. The preference list indicates the willingness of a person to choose other people as roommate. Then, a stable matching is defined as follows [126].

Definition 3 *In a stable matching, there is no such pair of people who are not matched as roommates, while both of them prefer each other to their current partners. In other words, there is no such a pair of people that both of them have a better choice than their current partners.*

By definition, a stable matching offers a desirable pairing strategy for a group of people. In our model, we regard each FUE in a femtocell as a person, who will pick another person in the same femtocell as roommate. The preference list of an FUE is determined by the level of interference caused by other FUEs. We then employ the effective algorithm proposed in [126] to solve the matching problem, and use the pairing result to select the duplex mode for the FUEs.

The proposed FUE paring strategy consists of three stages. In the first stage, each FUE proposes to and rejects other FUEs according to its preference list, and reduces its preference list by removing some undesired FUEs. In the second stage, the preference list of each FUE is further reduced, so that there is only one or no person in the list, which yields a stable matching solution. In the third stage, the matching solution is used by the FBS to determine the duplex mode selection strategy. The detailed procedure is presented in the following.

First Stage

First, each FUE establishes the preference list according to the interference power received from other FUEs. To implement this procedure, all FUEs send out pilot signals using a specific time slot. Then, the FUEs identify and measure the signal powers from other FUEs. Afterwards, each FUE inserts the other FUEs into the preference list in the descending order of received interference.

Initially, each FUE proposes to other FUEs following the order of the preference list. When an FUE i receives a proposal from another FUE j, the following strategy is adopted.

- FUE i rejects FUE j if it already holds a better proposal from another FUE.
- FUE i holds FUE j for consideration if FUE j is better than the one that it currently holds. Then, FUE i rejects the FUE that it currently holds.

An FUE stops to propose until a promise of consideration is received. If it receives a rejection, it will continue to propose to other FUEs following the order in its preference list. The propose and reject actions terminate when either of the following two conditions are satisfied.

- Every FUE holds a proposal.
- One FUE has been rejected by every other FUE.

In the second case, every FUE but the rejected FUE holds a proposal. This is because they all rejected him as they already have a better choice. Then, the following lemma can be applied to reduce the preference lists of the FUEs.

Lemma 28 *If FUE i rejects FUE j in the proposal sequence described above, then FUE i and FUE j cannot be partners in a stable matching.*

Proof: Among all the rejections that involve two FUEs who are partners in a stable matching M, there must be a rejection that happens at the *first time*, we denote it as FUE y rejects FUE x.

Since y rejects x, y must already held or later received a better proposal than x, denote as z. For stability of matching M, z must prefer its own partner w to y (otherwise both y and z prefer each other to their current partner). Since z proposed to y, he must be rejected by w, and this rejection must happen before z propose to y, so it is also before y rejects x. This contradicts to our assumption that y rejects x happens at the first time. ■

From Lemma 28, we derive the following useful corollary.

Corollary 1 *At any stage of the proposal process, if FUE i proposes to FUE j, we have that in a stable matching*
(i) FUE i cannot have a better partner than FUE j,
(ii) FUE j cannot have a worse partner than FUE i.

Proof: Since FUE i proposes to FUE j, it has been rejected by everyone better than FUE j. According to Lemma 28, FUE i cannot be partner with them. Thus part (i) is true.

Assume that FUE j has a worse partner FUE k. So FUE j prefers FUE i to FUE k. According to part (i), FUE i prefers FUE j to its own partner, which violates the definition of stable matching. Thus, the assumption does not hold. We thus conclude that part (ii) is also true. ■

According to Lemma 28, for the case that one FUE is rejected by everyone else, there is no stable matching exits since this FUE does not have a partner. To deal with this case, we set this FUE to work in the HD mode, i.e., it does not pair with any other FUEs for FD transmission. Then, the reduced problem can be solved as discussed earlier. With every FUE holds a proposal, the following corollary can be used to reduce the preference lists.

Corollary 2 *The preference list of FUE i, who holds a proposal from FUE j, can be reduced by deleting*

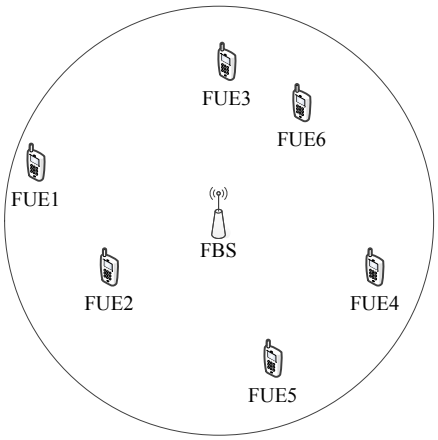

Figure 5.2: Example of preference lists for the six FUEs in a femtocell.

- *All those to whom FUE i prefers FUE j;*
- *All those who hold a proposal from a person whom they prefer to FUE i (including all those who have rejected FUE i).*

Proof: According to Corollary 1, the FUEs described above cannot be partners with FUE i in a stable matching. They can be removed from the preference list of FUE i. ■

In Fig. 5.2, we show an example of the preference lists in a femtocell with 6 FUEs. We have the following events.

1 proposes to 4, 4 holds 1,

2 proposes to 4, 4 rejects 2,

2 proposes to 6, 6 holds 2,

3 proposes to 5, 5 holds 3,

4 proposes to 1, 1 holds 4,

5 proposes to 3, 3 holds 5,

6 proposes to 1, 1 rejects 6,

6 proposes to 5, 5 rejects 6,

6 proposes to 2, 2 holds 6.

The preference lists can be reduced as

1: **4** ~~5 6 2 3~~

2: 4 **6** ~~3 5 1~~

3: **5** ~~4 2 1 6~~

4: **1** ~~2 3 6 5~~

5: **3** ~~1 6 2 4~~

6: ~~1 5~~ **2** ~~4 3~~

The purpose of reducing the preference lists is given in the following lemma.

Lemma 29 *If the preference list of every FUE contains just one FUE, then the lists specify a stable matching.*

Proof: Suppose that an FUE x prefers FUE y to the only FUE on its list. According to the propose strategy, x must be rejected by y, due to the fact that y has received a better proposal. However, the final proposal held by y is the only person on the reduced list of y, so y prefers this person to x (otherwise y would not reject x). Therefore, it is impossible that a pair of FUEs prefer each other to their own partners, which specifies a stable matching. ■

Although the preference lists are reduced in the first stage, some lists may still contain more than one person. This brings us to the second stage of the algorithm.

Second Stage

In the second stage, we further reduce the preference lists until each FUE holds only one proposal. The key is to find a cyclic sequence and initiate more rejections based on the sequence. We can prove that, with such rejections, a stable matching can be achieved.

An *all-or-nothing* cyclic sequence is defined as follows.

Definition 4 *Let* $\{a_1, \cdots, a_r\}$ *be a set of FUEs satisfying the following conditions.*

- *For* $i = 1, \cdots, r-1$*, the second person in* a_i*'s current reduced preference list is the first person in* a_{i+1}*'s, denote this person as* b_{i+1}*;*
- *The second person in* a_r*'s current reduced preference list is the first in* a_1*'s, denote this person as* b_1*.*

Algorithm 8: Find an All-or-Nothing Cyclic Sequence

do
  Let p_1 be an arbitrary FUE with preference list contains more than one FUE ;
  do
    $q_i \leftarrow$ the second FUE in p_i's current list ;
    $p_{i+1} \leftarrow$ the last person on q_i's current list (so that q_i is the first in p_{i+1}'s list) ;
  while (*the p sequence is not cyclic*);
while (*there is at least one FUE whose perference list contains more than one FUE*);
Denote p_s as the first element in the p sequence to be repeated, $r = s - 1$;
for $i = 1 : r$ **do**
  $a_i = p_{s+i-1}$;
end
Then $\{a_1, \cdots, a_r\}$ is an all-or-nothing sequence ;

Then, we adopt the algorithm presented in Algorithm 8 to find an all-or-nothing cyclic sequence. With the all-or-nothing sequence, we force each b_i to reject the proposal from a_i. Thus, each a_i turns to propose to b_{i+1}, the second favored FUE for a_i. With these rejections and proposals, all successors (those who rank after) of a_i can be deleted from the list of b_{i+1}, and b_{i+1} can be deleted from their lists. We continue the search of all-or-nothing circles and force rejections until each FUE holds only one proposal. However, whether these rejections cause instability to a stable matching remains uncertain. Next, we show that the rejections within an all-or-nothing circle would not cause instability to a matching that derived from reduced preference lists.

In a stable matching, if every FUE is partnered by someone on its reduced list, we say that such a matching is *contained* in the reduced lists. We have the following lemma.

Lemma 30 *Let $\{a_1, \cdots, a_r\}$ be an all-or-nothing circle, and b_i be the first person in a_i's reduced list, $1 \leq i \leq r$. Then, in any stable matching contained in these reduced lists, either a_i and b_i are partners for all values of i or for no value of i.*

Proof: Suppose for some i, a_i and b_i are partners in a stable matching that is contained in the reduced lists. Since a_i is last on b_i's reduced list, and b_i is second on a_{i-1}'s, a_{i-1} is must be present in b_i's reduced list. Thus, b_i prefers a_{i-1} to a_i. For stability, a_{i-1} must be partnered by someone he prefers to b_i, and the only such person in his reduced list is b_{i-1}. Repeating this argument shows that a_i and b_i must be partners for all values of i. ■

To show that the rejection within an all-or-nothing circle does not cause instability to a stable matching contained in the reduced lists, we need to consider the case that a_i and b_i are partners for all values of i in the circle.

Lemma 31 *Let $\mathcal{A} = \{a_1, \cdots, a_r\}$ and $\mathcal{B} = \{b_1, \cdots, b_r\}$. Suppose M is a stable matching contained in the reduced lists, with a_i and b_i being partners for all $1 \leq i \leq r$. Denote M' as the matching in which each a_i is partnered by b_{i+1}, and any FUE not in $\mathcal{A} \cup \mathcal{B}$ has the same partner as in M. Then, M' is stable.*

Proof: As shown in the proof for Lemma 30, each FUE in $\mathcal{B}$ obtains a better partner in M' than the one he had in M, while the each FUE in $\mathcal{A}$ gets a less favorable partner. Thus, the potential instability of M' must involve a_i. Suppose M' is not stable, since a_i prefers x, to b_{i+1} (his partner in M'), then there are three cases to consider.

1. a_i and x were partners in M (i.e., x is b_i). In this case, x prefers his new partner a_{i-1} to a_i, so there is no instability.

2. a_i prefers x to b_i. Since b_i is the first remaining choice for a_i, x is not in a_i's reduced list. Thus, x must has willingly rejected or has been forced to reject a_i. In the first case, x must have received a proposal from an FUE better than a_i. In the second case, x has received a better proposal due to the rejection. Therefore, x must prefer his partner in M' to a_i.

3. a_i prefers b_i to x. Hence, x is between b_i and b_{i+1} in a_i's original preference list. Since b_{i+1} is the second in a_i's reduced list, x must not be on a_i's reduced list. For the same argument in (b), x must prefer his partner in M' to a_i.

■

Lemma 31 demonstrates that the rejections and proposals introduced in Stage 2 maintain the stability of a matching contained in the reduced lists. Thus, we can find a stable matching adopting the procedure described earlier in Stage 2.

Third Stage

Based on the FUE matching result, we determine the FD/HD mode selection for the FUEs in Stage 3. Note that, the mode selection discussed in this part is regarded as an initial solution that does not consider the effect of inter-femtocell interference and cross-tier interference. The duplex modes may be refined due to inter-femtocell and cross-tier interferences in the greedy channel allocation algorithm described in Section 5.4.3.

For a pair of FUEs that are paired in the matching, denoted as FUE 1 and FUE 2, let H_{11} and H_{22} be the channel gains between the FBS and the two FUEs, respectively, and let H_{12} be the channel gain between the two FUEs. Denote p_1^u, p_2^u as the uplink powers, and p_1^d and p_2^d are the downlink powers of the two FUEs. The sum capacity of this pair of users with FD and HD modes can be derived as in (5.41) and (5.42), respectively.

$$\begin{aligned}\mathcal{C}(FD) = &\log_2\left(1+\frac{p_1^d H_{11}}{p_2^u H_{12}+N_0}\right)+\log_2\left(1+\frac{p_2^d H_{22}}{p_1^u H_{12}+N_0}\right)\\ &+\log_2\left(1+\frac{p_1^u H_{11}}{p_2^d \kappa+N_0}\right)+\log_2\left(1+\frac{p_2^u H_{22}}{p_1^d \kappa+N_0}\right)\end{aligned} \tag{5.41}$$

$$\begin{aligned}\mathcal{C}(HD) = \frac{1}{2}\Bigg\{&\log_2\left(1+\frac{p_1^d H_{11}}{N_0}\right)+\log_2\left(1+\frac{p_2^d H_{22}}{N_0}\right)\\ &+\log_2\left(1+\frac{p_1^u H_{11}}{N_0}\right)+\log_2\left(1+\frac{p_2^u H_{22}}{N_0}\right)\Bigg\}\end{aligned} \tag{5.42}$$

For each FUE pair, we compare the sum capacities of FD and HD as given in (5.41) and (5.42). We then select the duplex mode that achieves a higher capacity for the pair of FUEs.

5.4.2 Distributed Power Control Scheme

The power control problem under multiple channels can be extended from that in Section 5.3.2. In this part, we present the distributed implementation procedure for power control with the

SCALE algorithm. In particular, we focus on how to distributively obtain the solution that maximizes the lower bound (5.34).

Given the channel allocation and duplex mode selection results, the power control problem can be formulated as

$$\underset{\{p^u_{f,i}(n),p^d_{f,i}(n)\}}{\arg\max} \sum_f \sum_n \Big\{\log_2\left(1+\gamma^{FUE}_{f,i}(n)\right)+ \log_2\left(1+\gamma^{FBS}_{f,i}(n)\right)\Big\} \tag{5.43}$$

$$\text{subject to: } \sum_{n=1}^{N} p^u_{f,i}(n) \le P_{m1}, \ \ \forall f, \forall i \in \pi_f \tag{5.44}$$

$$\sum_{n=1}^{N} p^d_{f,i}(n) \le P_{m2}, \ \ \forall f, \forall i \in \pi_f \tag{5.45}$$

$$n \in \{1, \cdots, N\} \tag{5.46}$$

$$\{f,i\} \in \Pi_1 \cup \Pi_2 \cup \cdots \cup \Pi_n \cup \cdots \Pi_N, \tag{5.47}$$

where $\Pi_n = \{1, 2, \cdots, k, \cdots\}$ is the set of FBS-FUE links $\{f, i\}$ that operate on channel n.

Applying the approximation method in Section 5.3.2, we transform the original problem into the following problem that maximizes the lower bound of the objective function.

$$\underset{\{q_k(n)\}}{\arg\max} \sum_n \sum_{k\in\Pi_n} \{\alpha_k(n)\times \log_2\left(\frac{e^{q_k(n)}H_{kk}(n)}{\sum\limits_{l\in\Pi_n, l\neq k} e^{q_l(n)}H_{lk}(n)+N_0}\right)+\beta_k(n)\Big\} \tag{5.48}$$

$$\text{subject to:}$$

$$\sum_{n=1}^{N} e^{q_k(n)} \le P_{m1}, k \in \rho_n \tag{5.49}$$

$$\sum_{n=1}^{N} e^{q_k(n)} \le P_{m2}, k \in \theta_n, \tag{5.50}$$

where ρ_n is the set of FBS-FUE uplinks that operate on channel n, θ_n is the set of FBE-FUE downlinks that operate on channel n, H_{kk} is the channel gain of link k, and H_{lk} is the channel gain between the transceiver of link l and the receiver of link k.

The Lagrangian of the transformed problem is

$$L\left(\mathbf{q},\lambda,\mu\right)=\sum_{n}\sum_{k\in\Pi_n}\alpha_k\left(n\right)\log_2\left(\frac{e^{q_k(n)}H_{kk}\left(n\right)}{N_0+\sum\limits_{l\in\prod_n,l\neq k}e^{q_l(n)}H_{lk}\left(n\right)}\right)$$
$$+\beta_k\left(n\right)+\sum_{k\in\rho_n}\lambda_k\left(P_{m1}-\sum_{n=1}^{N}e^{q_k(n)}\right)+\sum_{k\in\theta_n}\mu_k\left(P_{m2}-\sum_{n=1}^{N}e^{q_k(n)}\right). \tag{5.51}$$

According to the first order necessary condition, we have that $\frac{\partial L(q,\lambda,\mu)}{\partial q_k(n)}=0$, for all $k\in\rho_n$, and $\frac{\partial L(q,\lambda,\mu)}{\partial q_k(n)}=0$, for all $k\in\theta_n$.

Thus, the following equations hold.

$$\begin{cases} p_k\left(n\right)=\frac{\alpha_k(n)}{\ln 2\lambda_k+\sum\limits_{l\in\prod_n,l\neq k}\frac{\alpha_l(n)H_{kl}(n)}{N_0+\sum\limits_{j\in\prod_n,j\neq l}p_j(n)H_{jl}(n)}},k\in\rho_n \\ p_k\left(n\right)=\frac{\alpha_k(n)}{\ln 2\mu_k+\sum\limits_{l\in\prod_n,l\neq k}\frac{\alpha_l(n)H_{kl}(n)}{N_0+\sum\limits_{j\in\prod_n,j\neq l}p_j(n)H_{jl}(n)}},k\in\theta_n. \end{cases} \tag{5.52}$$

In the dual domain, the dual variables λ_k and μ_k can be iteratively obtained through a gradient descent search.

$$\begin{cases} \lambda_k^{[t+1]}=\left[\lambda_k^{[t]}+\tau\left(\sum\limits_{n=1}^{N}p_k\left(n\right)-P_{m1}\right)\right]^+,k\in\rho_n \\ \mu_k^{[t+1]}=\left[\mu_k^{[t]}+\tau\left(\sum\limits_{n=1}^{N}p_k\left(n\right)-P_{m2}\right)\right]^+,k\in\theta_n, \end{cases} \tag{5.53}$$

where $[\cdot]^+=\max\left(0,\cdot\right)$, τ is the step size for each iteration, t is the index of iteration for the dual variables update. Thus, each FBS or FUE can locally update the dual variable λ_k or μ_k for link k.

Given the values of λ_k and μ_k, we obtain the solution for $p_k\left(n\right)$ using (5.52). Note that, these two equations are not closed-form expressions and $p_k\left(n\right)$ also appears at the right hand side (RHS). We propose a distributed algorithm to obtain the solution to problem (5.48).

Algorithm 9: Distributed Algorithm for Maximization of Lower Bound (5.48)

Initialize $t = 1$;
do
  Receive message $\sum_{l\neq k, l\in\Pi_n} \frac{\alpha_l(n)H_{lk}(n)}{N_0+\sum_{j\in\Pi_n, j\neq l} p_j(n)H_{jl}(n)}$ from MBS ;
  do
    Update $[p_k(1), ..., p_k(n), ...p_k(N)]$ as in (5.54) ;
    Update λ_k or μ_k as in (5.53) ;
  while (λ_k *or* μ_k *does not converge*);
  Send messages $\sum_{l\in\Pi_n, l\neq k} p_l(n) H_{lk}(n)$ and $H_{lk}(n)$ to MBS ;
  $t \leftarrow t + 1$;
while ($[p_k(1), ..., p_k(n), ...p_k(N)]$ *does not converge*);

Based on the expressions in (5.52), the power at iteration t can be updated as

$$\begin{cases} p_k^{[t+1]}(n) = \frac{\alpha_k(n)}{\ln 2\lambda_k^{[t]} + \sum_{l\neq k, l\in\Pi_n} \frac{\alpha_l(n)H_{kl}(n)}{N_0+\sum_{j\in\Pi_n, j\neq l} p_j(n)H_{jl}(n)}}, k \in \rho_n \\ p_k^{[t+1]}(n) = \frac{\alpha_k(n)}{\ln 2\mu_k^{[t]} + \sum_{l\neq k, l\in\Pi_n} \frac{\alpha_l(n)H_{kl}(n)}{N_0+\sum_{j\in\Pi_n, j\neq l} p_j(n)H_{jl}(n)}}, k \in \theta_n. \end{cases} \tag{5.54}$$

In (5.54), λ_k and μ_k can be locally updated by the FBS or FUE on link k. The term $\sum_{j\in\Pi_n, j\neq l} p_j(n) H_{jl}(n)$ is the interference received by the receiver of link l, which can also be measured locally as $\sum_{j\in\Pi_n, j\neq l} p_j(n) H_{jl}(n) = \frac{p_l(n)H_{ll}(n)}{SINR_l(n)}$. We assuming that $H_{lk}(n)$ can be locally estimated by link k through pilot signaling.

After measuring $\sum_{j\in\Pi_n, j\neq l} p_j(n) H_{jl}(n)$ and $H_{lk}(n)$, all the FBS's send these information to MBS. The MBS then calculates $\sum_{l\neq k, l\in\Pi_n} \frac{\alpha_l(n)H_{lk}(n)}{N_0+\sum_{j\in\Pi_n, j\neq l} p_j(n)H_{jl}(n)}$ for FBS-FUE link k, and sends the results back to each FBS. With the information from the MBS, each FBS or FUE can distributively update the Lagrangian multiplier and power. Repeating the process until converge, we obtain the solution to problem (5.48). The procedure of link k on channel n is described in Algorithm 9.

With the distributed power control solution that maximizes the lower bound (5.48), we then apply the SCALE algorithm described in Algorithm 7 to obtain the solution for the power control solution over multiple channels.

5.4.3 Greedy Channel Allocation Algorithm

Greedy Algorithm

Given the duplex mode selection results and power control solution, the channel allocation problem formulated in (5.12)–(5.19) becomes an integer programming problem. Solving it through exhaustive search incurs prohibitive high complexity.

In this section, we propose a greedy algorithm to allocate channels to FBS-FUE links. First, for each femtocell, we select the link with largest capacity, and denote $\{1, 2, \cdots, s, \cdots\}$ as the set of selected links. If link s is in the FD mode, we denote $s(u)$ and $s(d)$ as the paired uplink and downlink that compose the FD link s. Denote $\eta(n)$ as the set of FBS-FUE links that are allocated with channel n, $\varepsilon(n)$ as the set of links that are not allocated with channel n. Let $R(\eta(n) + s)$ and $R(\eta(n))$ be the sum rates of link sets $\eta(n) + s$ and $\eta(n)$ operate on channel n, respectively. Then, we define $\Delta(\eta(n) + s, \eta(n)) = R(\eta(n) + s) - R(\eta(n))$. Thus, $\Delta(\eta(n) + s, \eta(n))$ is the increment of objective value by allocating channel n to link s.

The procedure of the greedy channel allocation algorithm is given in Algorithm 10. For each channel, the FBS-FUE link that achieves the largest performance gain is chosen. If allocating the channel to such a link does not result in violation of the QoS requirements for all FUs and MUs, the channel is allocated to the link and we continue to search the link with the largest performance gain in the remaining links. If the allocation violates the QoS requirement of an MU or FU, we consider two cases depending on the duplex mode of the link. If the link is in the HD mode, the link cannot access the channel. If the link is in the FD mode, we first compare the performance gains of the two links that form the FD link, and forbidden the one with the lower gain to access the channel. Then, we update the SINRs of FUs and MUs. If the QoS requirements are satisfied, we move to the next round to find the link with the next largest capacity gain. If the QoS requirements are not satisfied, both links that form the FD link are not allowed to access the channel. This process terminates when all the links are examined or the allocation cannot achieve a positive performance gain.

Algorithm 10: Greedy Channel Allocation Algorithm

Initialize: $\eta(n) = \emptyset$, $\varepsilon(n) = \{1, 2, \cdots, s, \cdots\}$, $a_{s'}(n) = 0$, for all s', n ;
for $n = 1 : N$ **do**
  while $(\varepsilon(n) \neq \emptyset)$ & $(\Delta(\eta(n) + s, \eta(n)) > 0)$ **do**
    $s' \leftarrow \arg\max_{s \in \varepsilon(n)} \{\Delta(\eta(n) + s, \eta(n))\}$;
    Suppose $a_{s'}(n) = 1$;
    Update SINRs of MUs and FUs ;
    if (*the SINR requirements of MUs and FUs are satisfied*) **then**
      Set $a_{s'}(n) = 1$;
      $\eta(n) \leftarrow \eta(n) + s', \varepsilon(n) \leftarrow \varepsilon(n) - s'$;
    else
      if (s' *is in the FD mode*) **then**
        $s''_{\min} \leftarrow \arg\min_{\{s'(u), s'(d)\}} \{\Delta(\eta(n) + s'(u), \eta(n)), \Delta(\eta(n) + s'(d), \eta(n))\}$;
        $s''_{\max} \leftarrow \arg\max_{\{s'(u), s'(d)\}} \{\Delta(\eta(n) + s'(u), \eta(n)), \Delta(\eta(n) + s'(d), \eta(n))\}$;
        Set $a_{s''_{\min}}(n) = 0$;
        Update SINRs of MUs and FUs ;
        if (*the SINR requirements of MUs and FUs are satisfied*) **then**
          $\varepsilon(n) \leftarrow \varepsilon(n) - s', \varepsilon(n) \leftarrow \varepsilon(n) + s''_{\max}$;
        else
          Set $a_{s''_{\max}}(n) = 0$;
          $\varepsilon(n) \leftarrow \varepsilon(n) - s'$;
        end
      else
        $a_{s'}(n) = 0$;
        $\varepsilon(n) \leftarrow \varepsilon(n) - s'$;
      end
    end
  end
end

Performance Bound

We next derive a lower performance bound for the proposed greedy channel allocation algorithm. Suppose that the channel allocation algorithm takes L steps. Denote $e(l)$ as the lth link-channel pair chosen by the greedy algorithm. Let $\tau_l = \{e(1), e(2), \cdots, e(l)\}$ be the sequence of channel allocation. The increase of objective value with the l the link-channel allocation is given as

$$\Delta_l = \Delta(\tau_l, \tau_{l-1}) = R(\tau_l) - R(\tau_{l-1}). \tag{5.55}$$

Sum up Δ_l from $l = 1$ to $l = L$, we have

$$\begin{aligned}\sum_{l=1}^{L} \Delta_l &= R(\tau_L) - R(\tau_{L-1}) + \cdots + R(\tau_1) - R(\tau_0) \\ &= R(\tau_L) - R(\tau_0) = R(\tau_L).\end{aligned}$$

Let G_I be the interference graph, where each vertex represents a link. When allocating a channel to an FUE–FBS link results in the violation of the SINR requirement of another FUE–FBS link, there will be an edge between these two links in the graph, indicating that they cannot simultaneously utilize the same channel due to inter-femtocell interference. For two link-channel pairs $e(l)$ and $e(l')$, we say $e(l)$ *conflicts* with $e(l')$ when there is an edge between the two links. Denote Φ as the optimal solution of channel allocation, we define φ_l as the subset of Φ that conflicts with allocation $e(l)$ but not with the previous allocations $\{e(1), e(2), \cdots, e(l-1)\}$.

For two feasible allocations τ_1 and τ_2, we show that their performance difference $\Delta(\tau_2, \tau_1) = R(\tau_2) - R(\tau_1)$ has the following properties.

Property 4 *For feasible link-channel pairs ω_1, ω_2 and χ, if $\omega_1 \subseteq \omega_2$ and $\chi \cap \omega_2 = \emptyset$, then* $\Delta(\omega_2 \cup \chi, \omega_1 \cup \chi) \leq \Delta(\omega_2, \omega_1)$.

Property 5 *For feasible link-channel pairs ω, χ_1 and χ_2, if $\omega \cap \chi_1 = \emptyset$, $\omega \cap \chi_2 = \emptyset$ and $\chi_1 \cap \chi_2 = \emptyset$, then* $\Delta(\varphi \cup \chi_1 \cup \chi_2, \omega) \leq \Delta(\omega \cup \chi_1, \omega) + \Delta(\omega \cup \chi_2, \omega)$

In Property 4, considering the left hand side (LHS) of the inequality, the link-channel pairs in $\omega_2 - \omega_1$ receive additional interference from the links in χ, resulting in a smaller increase in the objective value. In property 2, considering the LHS of the inequality, the links in χ_1 and χ_2 interfere with each other, resulting in a smaller increase in the objective value than the case on the RHS.

Since the link-channel pair with the maximum performance improvement is chosen in each step of the greedy algorithm, for any $\chi \in \varphi_l$, we have

$$\Delta(\tau_{l-1} \cup \chi, \tau_{l-1}) \leq \Delta_l. \tag{5.56}$$

Suppose the greedy algorithm stops in L steps, the following lemma can be derived

Lemma 32 $R(\Phi) \leq R(\tau_L) + \sum_{l=1}^{L} \sum_{\chi \in \varphi_l} \Delta(\tau_{l-1} \cup \chi, \tau_{l-1})$.

Proof: Based on the properties of $\Delta(\cdot,\cdot)$ function, we have the following inequalities.

$$\begin{aligned} & R\left(\left(\cup_{i=l+1}^{L}\varphi_i\right) \cup \tau_l\right) = R\left(\left(\cup_{i=l+2}^{L}\varphi_i\right) \cup \tau_l\right) + \Delta\left(\left(\cup_{i=l+1}^{L}\varphi_i\right) \cup \tau_l, \left(\cup_{i=l+2}^{L}\varphi_i\right) \cup \tau_l\right) \\ & \overset{(1)}{\leq} R\left(\left(\cup_{i=l+2}^{L}\varphi_i\right) \cup \tau_l\right) + \Delta(\varphi_{l+1} \cup \tau_l, \tau_l) \overset{(2)}{\leq} R\left(\left(\cup_{i=l+2}^{L}\varphi_i\right) \cup \tau_{l+1}\right) + \Delta(\varphi_{l+1} \cup \tau_l, \tau_l) \\ & \overset{(3)}{\leq} R\left(\left(\cup_{i=l+2}^{L}\varphi_i\right) \cup \tau_{l+1}\right) + \sum_{\chi \in \varphi_{l+1}} \Delta(\chi \cup \tau_l, \tau_l). \end{aligned} \tag{5.57}$$

Inequality (1) results from Property 4; Inequality (2) is because the $(l+1)$th allocation increases the objective value; Inequality (3) is an application of Property 5.

Rewrite (5.57) as

$$R\left(\left(\cup_{i=l+1}^{L}\varphi_i\right) \cup \tau_l\right) - R\left(\left(\cup_{i=l+2}^{L}\varphi_i\right) \cup \tau_{l+1}\right) \leq \sum_{\chi \in \varphi_{l+1}} \Delta(\chi \cup \tau_l, \tau_l). \tag{5.58}$$

By definition, φ_{L+1} does not conflict with the previous L allocations, indicating that the greedy algorithm takes at least $L+1$ steps, which is a contradiction. Thus, $\varphi_{L+1} = \emptyset$. It is obvious that $\tau_0 = \emptyset$. Then, with induction on (5.58) from $l = 0$ to $l = L-1$, we have $R(\Phi) \leq R(\tau_L) + \sum_{l=1}^{L} \sum_{\chi \in \varphi_l} \Delta(\tau_{l-1} \cup \chi, \tau_{l-1})$. ■

Lemma 33 *The maximum size of φ_l is equal to the degree of the link selected in step l in the interference graph G_I, denoted as $D(l)$.*

Proof: By definition, φ_l conflicts with the link-channel allocation selected by $e(l)$, the number of allocations that possibly conflict with $e(l)$ is the degree of the node that represent the selected link in the interference graph. ■

Theorem 7 *The solution solved by the greedy algorithm is at least $1/(1+D_{\max})$ of the global optimum, where $D_{\max}$ is the maximum node degree in the interference graph G_I.*

Proof: Applying Lemma 32, Lemma 33 and (5.56), we have

$$R\left(\Phi\right) \leq R\left(\tau_L\right) + \sum_{l=1}^{L} D\left(l\right)\Delta_l = R\left(\tau_L\right) + \bar{D}\sum_{l=1}^{L}\Delta_l = \left(1+\bar{D}\right)R\left(\tau_L\right),$$

where $\bar{D} = \sum_{l=1}^{L} D\left(l\right)\Delta_l / \sum_{l=1}^{L}\Delta_l$. Since $D\left(l\right) \leq D_{\max}$, we have $\bar{D} \leq \sum_{l=1}^{L} D_{\max}\Delta_l / \sum_{l=1}^{L}\Delta_l = D_{\max}$. Thus, the lower bound of $R\left(\tau_L\right)$ is derived and we have

$$\frac{1}{1+D_{\max}} R\left(\Phi\right) \leq R\left(\tau_L\right) \leq R\left(\Phi\right). \tag{5.59}$$

■

5.4.4 Convergence Analysis

In this section, we prove that the proposed iterative framework converges to a final solution.

Theorem 8 *The proposed iterative framework for joint duplex mode selection, power control, and channel allocation converges to a final solution.*

Theorem 9 *The number of iterations required to converge is upper bounded by* $U = \max\left\{U_1, U_2\right\}$, *where*

$$\begin{cases} U_1 = \max\limits_{n}\left\{F - \left\lceil \frac{p_b H_{b,m}(n) - N_0\Gamma_1}{\Gamma_1 \cdot \max\left\{P_{m_1}, P_{m_2}\right\} \cdot \sum\limits_{s=1}^{F} H_{sm}(n)} \right\rceil\right\} \\ U_2 = \max\limits_{n}\left\{F - 1 - \left\lceil \frac{p_s^{[1]}(n) H_{ss}(n) - N_0\Gamma_2 - I_{pf}(n)\Gamma_2}{\Gamma_2 \cdot \max\left\{P_{m_1}, P_{m_2}\right\} \cdot \sum\limits_{r=1, r\neq s}^{F} H_{rs}(n)} \right\rceil\right\}, \end{cases} \tag{5.60}$$

where $H_{sm}(n)$ *is the channel gain between the transmitter of link* s *to the MUE,* $H_{rs}(n)$ *is the channel gain between the transmitter of link* r *and the receiver of link* s, $p_s^{[1]}(n)$ *is the transmit power of link* s *at the first iteration.*

Proof: In the greedy channel allocation algorithm, to guarantee that the SINRs of MUs and FUs operate on channel n are no less than the predefined threshold, some FBS-FUE links will be forbidden to access channel n.

We first consider the MU that uses channel n. Without loss of generality, suppose channel n is used for downlink transmission. Then, the SINR of the MUE should satisfy

$$\gamma_{MUE}(n) = \frac{p_b H_{b,m}(n)}{N_0 + \sum_{s=1}^{F} a_s(n) p_s(n) H_{sm}(n)} \geq \Gamma_1, \tag{5.61}$$

which can be re-written as

$$\sum_{s=1}^{F} a_s(n) p_s(n) H_{sm}(n) \leq \frac{p_b H_{b,m}(n)}{\Gamma_1} - N_0. \tag{5.62}$$

It follows the power constraints that

$$p_s(n) \leq \max\left\{P_{m_1}, P_{m_2}\right\}. \tag{5.63}$$

Also note that the inequality

$$\sum_{s=1}^{F} a_s(n) H_{sm}(n) \leq \sum_{s=1}^{F} a_s(n) \sum_{s=1}^{F} H_{sm}(n) \tag{5.64}$$

holds. Applying these two inequalities, we have

$$\sum_{s=1}^{F} a_s(n) p_s(n) H_{sm}(n) \leq \max\left\{P_{m_1}, P_{m_2}\right\} \sum_{s=1}^{F} H_{sm}(n) \sum_{s=1}^{F} a_s(n). \tag{5.65}$$

If the following equality (5.66) is satisfied, (5.62) must be satisfied.

$$\max\left\{P_{m_1}, P_{m_2}\right\} \sum_{s=1}^{F} H_{sm}(n) \sum_{s=1}^{F} a_s(n) \leq \frac{p_b H_{b,m}(n)}{\Gamma_1} - N_0. \tag{5.66}$$

Inequality (5.66) shows that when $\sum_{s=1}^{F} a_s(n)$ equals to any integer less than $\frac{p_b H_{b,m}(n) - N_0\Gamma_1}{\Gamma_1 \cdot \max\left\{P_{m_1}, P_{m_2}\right\} \cdot \sum_{s=1}^{F} H_{sm}(n)}$, the SINR constraint (5.62) must be satisfied. Thus, the largest feasible value of $\sum_{s=1}^{F} a_s(n)$ is

$$\sum_{s=1}^{F} a_s(n) = \left\lceil \frac{p_b H_{b,m}(n) - N_0\Gamma_1}{\Gamma_1 \cdot \max\left\{P_{m_1}, P_{m_2}\right\} \cdot \sum_{s=1}^{F} H_{sm}(n)} \right\rceil. \tag{5.67}$$

Eqn. (5.67) reveals that when the number of links operate on channel n is equal to $\frac{p_b H_{b,m}(n) - N_0\Gamma_1}{\Gamma_1 \cdot \max\left\{P_{m_1}, P_{m_2}\right\} \cdot \sum_{s=1}^{F} H_{sm}(n)}$, the SINR requirement of the MU is always satisfied. Thus, the number of links that are forbidden to access channel n is no larger than:

$$\Theta(n) = F - \left\lceil \frac{p_b H_{b,m}(n) - N_0\Gamma_1}{\Gamma_1 \cdot \max\left\{P_{m_1}, P_{m_2}\right\} \cdot \sum_{s=1}^{F} H_{sm}(n)} \right\rceil. \tag{5.68}$$

Thus, by forbidding at most $\Theta(n)$ links to access channel n, the SINR requirement of the MU using channel n is always satisfied. Therefore, with the greedy channel allocation algorithm, the links chosen to access channel n would remain the same and the solution converges. To forbid $\Theta(n)$ links to access channel n, we need to execute the algorithm at most $\Theta(n)$ times. Consider all the N channels, the number of iterations required to converge is given by $U_1 = \max_n \{\Theta(n)\}$.

We next consider the sth FBS-FUE link operating on channel n. To meet the SINR requirement, we have

$$\sum_{r=1, r\neq s}^{F} a_r(n) p_r(n) H_{rs}(n) \leq \frac{p_s(n) H_{ss}(n)}{\Gamma_2} - N_0 - I_{pf}(n). \tag{5.69}$$

According to the greedy channel allocation algorithm, the number of channels operating on channel n is non-increasing at each iteration (i.e., the inter-femtocell interference on channel n

is non-increasing). Then, from (5.54), $p_s(n)$ is non-decreasing at each iteration, and $p_s(n) \geq p_s^{[0]}(n)$.

Recall the other two inequalities (5.63) and (5.64). If the following inequality (5.70) is satisfied, the SINR constraint (5.69) must hold.

$$\begin{aligned} &\max\{P_{m1}, P_{m2}\} \sum_{r=1, r\neq s}^{F} H_{rs}(n) \sum_{r=1, r\neq s}^{F} a_r(n) \\ &\leq \frac{p_s^{[0]}(n) H_{ss}(n)}{\Gamma_2} - N_0 - I_{pf}(n). \end{aligned} \tag{5.70}$$

Adopting the similar procedures described earlier, the number of iterations required to converge is thus given by $U_2 = \max_n \{\Xi(n)\}$, where

$$\Xi(n) = F - 1 - \left\lceil \frac{p_s^{[0]}(n) H_{ss}(n) - \Gamma_2 \left(N_0 + I_{pf}(n)\right)}{\Gamma_2 \max\{P_{m1}, P_{m2}\} \sum_{r=1, r\neq s}^{F} H_{rs}(n)} \right\rceil.$$

Since the SINR requirements of both MUs and FUs should be satisfied, the final upper bound of number of required iterations is $U = \max\{U_1, U_2\}$. ■

5.5 Simulation Study

In this section, we validate the performances of the proposed duplex mode selection (DMS) strategy, power control (PC), and channel allocation (CA) algorithms with Matlab simulations. We consider a macrocell overlaid with multiple femtocells, as shown in Fig. 5.1, and evaluate the sum rate of all the femtocells. The radius of the macrocell is 500 m. We consider a total bandwidth of 4 MHz that are divided into 160 channels. The transmit power of the MBS is set as 35 dBm, while the transmit power of the MUEs has five levels ranging from 10 dBm to 30 dBm according to the distance between the MUE and MBS. The power budget of an FBS and an FUE is set to 30 dBm and 25 dBm, respectively. The noise power spectrum density is assumed to be -174 dBm/Hz. We employ the ITU path loss model for both indoor and outdoor environments [59]. All channels experience Rayleigh block fading.

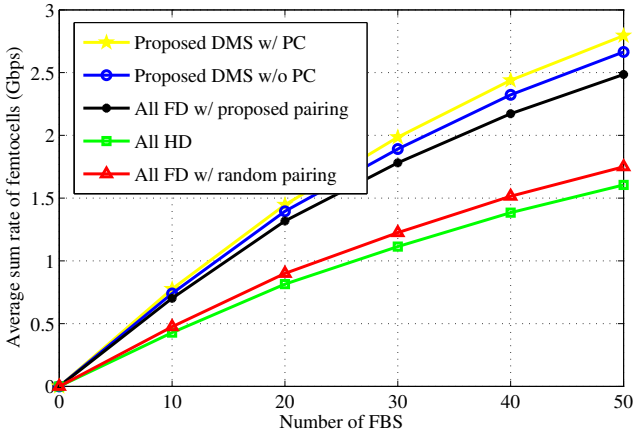

Figure 5.3: Average sum capacity versus the number of FBS under different duplex modes. The average number of FUE is five, the radius of a femtocell is 20 m, and κ=0.1.

We first present the performance of the proposed DMS strategy and PC algorithm in Figs. 5.3–5.6. In Fig. 5.3, we compare the performances of five schemes: (i) Applying both the proposed DMS strategy and PC algorithm, (ii) proposed DMS strategy without PC, (iii) all FD with the proposed pairing strategy without PC, (iv) all FD with a random pairing strategy without PC, and (v) all HD without PC. In scheme (iii), all pairs of FUEs that are matched with the proposed matching strategy adopt FD transmission. In scheme (iv), FUEs are randomly paired, and all the paired FUEs adopt FD transmission. In scheme (v), all FUEs adopt HD transmission. Comparing the performance of schemes (i) and (ii), a considerable gain can be achieved with the proposed PC algorithm. This is because the power of each FBS and FUE is set to a proper value that both improves data rate and mitigates inter-femtocell interference. It can be observed that the FD transmission achieves higher sum rates than HD transmission due to improved spectrum utilization. With the proposed DMS strategy, scheme (ii) achieves better performance than the all-FD and all-HD schemes, since we dynamically adjust the duplex mode for each pair of FUEs by choosing the mode that offers a higher rate. It can also be observed that the pairing strategy based on stable roommate matching outperforms the random pairing strategy, indicating that the stable matching provides a relatively better solution by pairing the FUEs with small interference to each other.

In Fig. 5.4, we present the performances under different numbers of FUEs. With the same reasons discussed above, the proposed DMS and PC schemes outperform the other schemes.

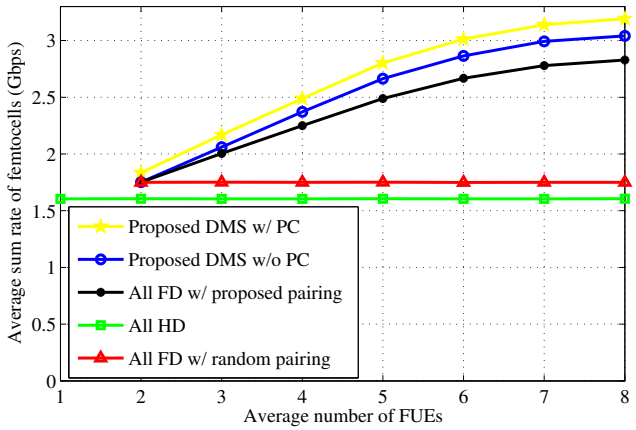

Figure 5.4: Average sum capacity versus the average number of FUEs under different duplex modes. The number of FBS is 50, the radius of a femtocell is 20 m, and κ=0.1.

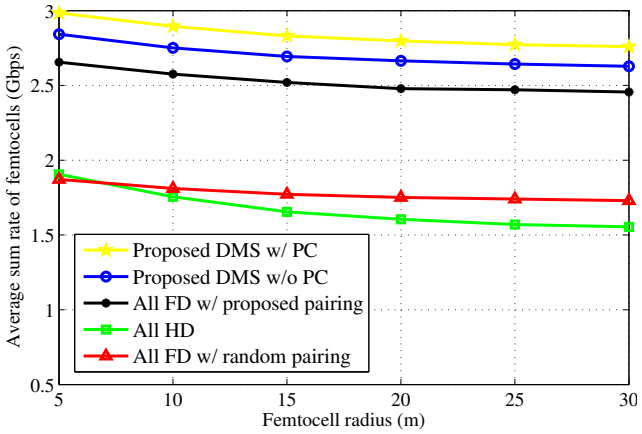

Figure 5.5: Average sum capacity versus the radius of a femtocell. The number of FBS is 50, the average number of FUE is five, and κ=0.1.

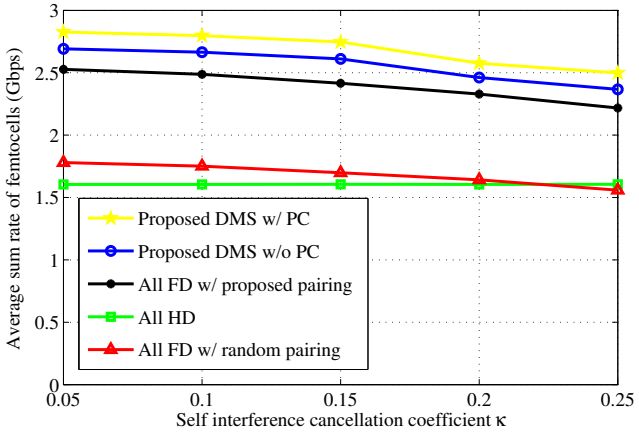

Figure 5.6: Average sum capacity versus the self interference cancellation coefficient κ. The number of FBS is 50, the average number of FUE is five, and the radius of a femtocell is 20 m.

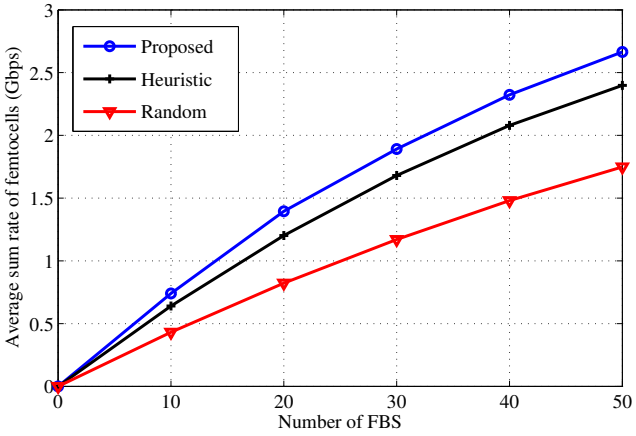

Figure 5.7: Average sum capacity versus the number of FBS's under different channel allocation schemes. The radius of a femtocell is 20 m, and κ=0.1.

With random pairing strategy, the performance of FD transmission is only slightly better than that of HD transmission (as in Fig. 5.3), since the intra-femtocell interference degrades the data rates of FUEs. The FD transmission with proposed pairing strategy achieves better performance than the one with random pairing, since the intra-femtocell interference is mitigated by properly pairing the FUEs for FD transmission. As the number of FUE increases, the gain of the proposed pairing strategy becomes more significant, since more pairs of FUEs benefit from the reduced intra-femtocell interference.

In Fig. 5.5, we evaluate the performances under different radii of a femtocell. As the radius of a femtocell increases, the average distance between an FUE and an FBS also increases, resulting in the performance degradations of all schemes. With random pairing, the performance of FD transmission is even worse than the HD transmission when the radius of a femtocell is small, i.e., the size of cell is small. When the size of a femtocell is small, FUEs are close to each other, the intra-femtocell interference becomes the dominating factor that impacts the SINR. Thus, the FD transmission with random pairing achieves worse performance than HD transmission. With the proposed FUE pairing and DMS strategy, the performance can be significantly improved, the advantage of FD over HD can be maintained.

In Fig. 5.6, we evaluate impact of κ. With a large κ, the performance of FD transmission with random pairing can be worse than that of HD. This shows that when the self interference

cancellation cannot be well performed, FD transmission losses its advantage over HD transmission. With the proposed pairing, DMS, and PC schemes, the performances of FD transmissions are improved as discussed before.

Fig. 5.7 compares the performance of different channel allocation schemes. We employ a heuristic algorithm with the idea proposed in [128] as a benchmark. In this heuristic algorithm, the FBS-FUE link that causes the smallest interference to the MUE or MBS on a channel is firstly chosen to access to the channel, and such process is continued until the QoS requirement of an MU or FU is violated. For the random allocation scheme, we randomly choose an FBS-FUE link at each step until the QoS requirement of an MU or FU is violated. It can be seen that the proposed scheme outperforms the other two schemes, since the links with higher channel gains are always firstly chosen, and whenever the QoS requirements are not satisfied, we continue to search for the link with the maximal gain among the other possible links.

5.6 Related Work

This work is related to many prior works on CR networks. For example, see [128, 131, 132]. In [131], a power control problem was formulated to maximize the sum rate of SUs, subject to constraints on the interference power caused to PUs. Since the objective and constraints can be transformed into posynomials, the problem can be modeled as a geometric programming and solved with existing methods. With the same constraint for protecting PU transmissions, the objective of [128] is to maximize the number of SUs that can be supported, and the problem was formulated as a mixed integer linear programming (MILP). Due to the prohibitive high complexity of MILP, a heuristic algorithm was proposed, in which channels would first be allocated to the secondary base station (SBS) with less interference to PU, and the SBS would first allocate channels to the SU with least channel gain. This way, the allocation that contributes less to the interference would first be selected, so that the secondary BS could serve more SUs. In [132], a two stage power and channel allocation scheme was proposed to maximize the throughput of SUs and guaranteed the SINR requirements of all the PUs and SUs. In the first stage, the SBS dynamically adjust the transmission power on each channel to control the interference and guarantee the SINRs of PUs. In the second stage, the maximal bipartite

matching algorithm is applied for SBS to allocate channels to SUs. Compared to these works, we consider a FD-empowered femtocell network with more complicated interference scenario.

The application of FD in wireless communication has drawn lots of interests, with recent advances reported in [122, 133–137]. Most of these works consider the scenario of relay assisted communication since a relay network directly benefit from the FD operations of relay nodes. In [133], a cooperative cognitive radio network was considered where SUs relay the signals of PUs with the FD pattern. Taking the residual self interference into account, this paper aims to maximize the sum rate of cognitive users and reduce the self-interference. The outage probability of FD relaying cognitive radio network was analyzed in [134], and the optimal power allocation strategy that minimizes the outage probability of SUs was proposed. In [135], the authors considered relay selection in a FD amplify-and-forward cooperative communication with the objective to maximize the channel capacity. In [136], the concept of effective capacity was introduced as the QoS provisioning measurement. Then, dynamic resource allocation schemes were proposed to maximize the effective capacity, under both amply-forward and decode-forward scenarios. Another perspective to improve the performance to relay network is to dynamically switch between half duplex (HD) and FD [137]. The authors analyzed the spectral efficiencies of both modes and proposed a power control scheme to optimize the performance. In [122], the capacities of FD transmissions with different interference suppression techniques were derived, then the sum rate maximization problem was formulated and solved with efficient algorithms.

5.7 Conclusion

In this chapter, we investigated the problem of joint duplex mode selection, channel allocation, and power control for FDCFNs. We employed stable roommate matching to model the FUE pairing problem, and proposed a duplex selection strategy based on the pairing result. We also developed a distributed power control algorithm and a greedy algorithm for channel allocation with proven performance bound. The convergence of the proposed iterative framework was proved. Simulation evaluations show that the proposed schemes outperform several other benchmark schemes with considerable capacity gains.

Chapter 6

Providing High-Data-Rate Service to Large Number of Users with BS Cooperation

6.1 Introduction

The recent wide development of user terminals, e.g., smartphones and tablets, has triggered a drastically increasing demand for high data rate services. With limited spectrum, such demands necessitate more efficient use of the spectrum resource to improve the capacity of wireless networks. To this end, small cells have been recognized as an effective means of enhancing network capacity [138, 139]. Compared to a single, macrocell base station (BS) with high power and large coverage area, a small cell network (SCN) consists of multiple low power and spatially separated small-scale BS's with small coverage areas. With such an architecture, the distance between the transmitter and receiver is greatly reduced, resulting in high signal to interference plus noise ratio (SINR). The low power of each BS enables more efficient spatial reuse of spectrum, which in turn improves the capacity of the entire network. Moreover, compared to traditional macrocell BS that brings high leasing and maintenance fees, SCN can greatly reduce the cost of wireless operators.

However, the benefits of SCN come at the price of a more complicated network architecture. The following features pose great challenges to the design and operation of an SCN.

- *A large number of small cell BS's*: The number of SCN BS's within a given area is expected to be much larger than that of macrocell BS's. With limited backhaul capacity, it would be difficult to perform centralized control and coordination.

- *Vulnerability to interference:* Due to the small coverage area, small cell BS's may be close to each other in a hotspot, which may easily cause strong inter-cell interference. If

different bands are assigned for neighboring cells, the available bandwidth for each cell will be greatly reduced. Moreover, deploying a small cell tier over an existing cellular network may also cause inter-tier interference if the two tiers occupy the same spectrum band.

- *Irregular coverage area*: Unlike traditional cellular networks with a hexagonal coverage area for each BS, the coverage area of a small cell BS is usually irregular. To mitigate interference as well as guaranteeing network connectivity, it is expected that the overlapping coverage areas of different small cell BS's are small and there are no coverage holes. These requirements make the deployment of small cell BS's more complicated.

- *Limited power budget*: Compared to a macrocell BS, the power of a small cell BS is quite limited. It is thus critical to efficiently utilize the power resource. Advanced technologies are needed to deal with the case when a large of users are served by a small cell BS under a stringent power constraint.

With increasing number of mobile users, these problems of SCN would be aggravated in hotspots with a heavy traffic load. When large amount of users assemble in a certain area (e.g., a shopping mall or a football stadium), the small cell BS's could be overloaded and unable to serve each user with limited spectrum resources. Under this circumstance, traffic congestion and call outage may frequently happen, resulting in degradation of user quality of service (QoS). Although SCN is initially regarded as a solution to guarantee user QoS in hotspots, supporting high data rate services to a large number of users now becomes a new challenge.

In this chapter, we focus on the problem of exploiting cooperative small cells to provide high data rate services for hotspots. Since the SCN needs to provide high data rate services to large number of users, it faces many new challenges when applied in hotspots. Although the proposed schemes in prior work [140–142] can deal with the overloading problem in hotspots, they are based on the traditional cellular network architecture and the traffic loads mainly consist of voice services with low requirements for data rate. Therefore, these schemes cannot be directly applied in the SCN to serve hotspots. Furthermore, the inherent features of SCN

as mentioned before introduce new challenges in the design of SCNs. Particularly, the large number of SCN BS's makes it a great challenge to adopt centralized network control strategies. Thus, decentralized strategies that allow local cooperation among the small cell BS's are highly appealing. To overcome these problems and satisfy the high capacity requirement at hotspots, we propose a cooperative small cell networks (CSCN) architecture in this chapter. Based on the SCN architecture, the proposed CSCN leverages several existing technologies with flexible designs to improve the spectrum utilization and network capacity, so as to provide enhanced QoS to users in the hotspots. The goal of this chapter is to provide an insightful look into the architecture and achievable performance enhancements of the CSCN. An illustrative example and a simulation study are presented to demonstrate the high potential of the proposed CSCN approach.

The remainder of this chapter is organized as follows. We first review several existing technologies for capacity improvement and serving hotspots. The concept of CSCN is presented next, followed by detailed discussions on various technical aspects of the proposed architecture. An illustrative example is provided along with a simulation study to demonstrate the performance of CSCN. We then discuss open problems and conclude this chapter.

6.2 Overview of Existing Technologies

6.2.1 Capacity Enhancement

With the idea of reducing transmission distance, a major approach to improve capacity can be classified as deploying access points that are close to users. The main techniques include Wi-Fi access points, femtocells [143], and distributed antenna systems [144]. With high-speed wireline connections and wireless routers as access points, the Wi-Fi technology has been widely adopted due to its salient features of low cost and easy deployment. However, it operates on unlicensed spectrum bands, and adopts contention-based MAC protocols, making it hard to guarantee the QoS requirements for a large number of users.

The femtocell concept was proposed for QoS provisioning and capacity enhancement. Femtocells are user deployed indoor low power access points that operate on licensed spectrum

band with typical coverage range of $10 - 50$ m. The small propagation loss and low power enable high data rates and spectrum reuse, benefiting both indoor users and wireless operators. Furthermore, the traffic burden of cellular network is reduced since indoor users are served by femtocells. However, the femtocell technology only improves indoor coverage, while adding femtocells over existing a cellular network potentially causes interference to outdoor cellular users.

The drawbacks of femtocell network are largely caused by the chaotic nature of its deployment by users. Effective control and coordination are necessary for overcoming such drawbacks. To this end, operator-deployed Distributed Antennas System (DAS) offers an effective solution and has been applied to improve both indoor and outdoor capacity. The basic idea of DAS is to deploy spatially distributed antenna units that are close to users, and these antenna units are connected to BS's with dedicated wireline connections. As femtocells, DAS brings about benefits on improved link quality and spatial diversity. However, a potential problem for DAS may be the limited capability of antenna units. Since the antenna units can only act as transceivers, different antenna units cannot distributively coordinate/cooperate with each other. Thus, the BS's need to perform a centralized control over all the connected antenna units, which brings new challenges to the processing units in BS and the backhaul, especially when the number of antenna units is large. Besides DAS, relay was proposed as another approach to improve coverage and capacity. In areas with poor coverage, relay nodes are deployed to receive, decode or amplify, and forward the signals for users, resulting in improved SINR and network capacity [145]. However, since the relay nodes only serve as transceivers without the coordination capability, a BS needs to perform centralized control for the resource allocations of all the relay nodes [146]. Thus, relay enhanced cellular networks bear similar limitations as DAS.

The essence of the above capacity improvement techniques is to improve SINR and spectral efficiency. According to Shannon's formula, increase the total bandwidth is most effective to improve network capacity. However, given the limited spectrum resource, the only choice is to create spectrum reuse opportunities to improve spectrum utilization. From the spatial domain, directional antennas, Multiple-Input and Multiple Output (MIMO), and smart antennas

all exploit spatial diversity to achieve spatial multiplexing gain. With proper design of antenna parameters, the application of directional antennas can reduce the undesired signal leakage, thus creating more spatial reuse opportunities in wireless networks [147]. For a MIMO system, through adjusting beamforming parameters according to the channel conditions of different antennas, the capacity can also be effectively improved. For a smart antennas system, a BS first estimates the location of a user based on the arrival direction of uplink signals. It then controls the downlink transmission beam pattern so that the desired user is served while other users are not interfered. This way, it is unnecessary to assign different channels or time slots to users served by the same BS; each user can be allocated with more resource, resulting in higher network capacity.

Furthermore, the cognitive radio (CR) technology [148] allows secondary users to opportunistically access the channels that are currently not occupied by primary users. With this approach, the spectrum opportunities can be utilized by the secondary users, thus higher spectrum efficiency is achievable. Nevertheless, due to limited spectrum sensing capability, the CR system faces several problems such as mis-detection, false-alarm, and hidden-terminals. Consequently, existing CR systems, e.g., IEEE 802.22 Wireless Regional Area Networks (WRAN), largely depend on a spectrum database to acquire channel state information.

6.2.2 Serving Hotspots

To satisfy the high capacity requirements at hotspots, several techniques were proposed based on the cellular network architecture, including cell splitting [149], cell sectoring [140], channel borrowing [141], and load balancing [142]. The cell splitting method is essentially similar to the small cells approach, where the original cells are split into smaller cells, so that the average number of available channels within a given area is increased. The cell sectoring technique employs directional antennas at each BS and each BS serves multiple sectors with different spectrum bands, thus reducing the co-channel interference among different cells and improving the frequency reuse efficiency.

Through channel borrowing, a cell borrows channels from adjacent cells to serve the users in a hotspot. Last but not least, load balancing provides another approach to deal with the heavy

traffic load in hotspots. When a cell is overloaded due to the hotspots in its coverage area, it can handover some of its users to adjacent cells. This way, the traffic burden in the hotspot cell can be mitigated, and the QoS of its users can be guaranteed.

6.3 Concept of Cooperative Small Cell Networks

CSCN is an extension of SCN for enabling local cooperation among neighboring small cell BS's. It consists of multiple small BS's that are equipped with directional antennas, and are connected to each other via a wireline backhaul through the X2 interface. The BS deployment pattern and BS coordination are carefully designed. The objective of CSCN is to improve the user QoS and the total network capacity in hotspots. The key components of CSCN are BS deployment, BS coordination based dynamic resource management, and interference coordination.

- *Base station deployment*: In the CSCN, the BS's are deployed according to the environment and average demand (with certain amount of redundancy). Then, the coverage area of each BS is adjusted to an appropriate shape to improve spatial reuse and reduce interference, which can be realized by applying directional antennas with adjustable parameters. With the development of hardware technology, BS's can be produced with much smaller size than before, and the coverage area of each BS can be flexibly controlled [150]. Therefore, flexible coverage areas and deployment patterns of antennas are feasible in the CSCN. When CSCN BS's are deployed in a hotspot, the environment information, such as the architectural layout, is utilized to determine the optimal shapes of the coverage areas.

- *Dynamic resource management:* In hotspots, the user distribution and traffic load may change quickly over time. Therefore, the CSCN should dynamically allocate spectrum resources by jointly considering the instantaneous requirements and interference mitigation, in order to improve capacity and accommodate the load. We propose a decentralized BS cooperation scheme that enables channel borrowing and user handover among neighboring BS's.

- *Interference coordination*: In the CSCN, BS's can sense the spectrum occupation condition and exchange information with neighboring BS's or the cellular BS's. They then coordinate with each other to mitigate interference as well as efficiently utilizing spectrum resources. Since the transmission beam of a mobile device (MD) is omni-directional, MDs served by neighboring CSCN BS's may cause uplink interference to each other. Moreover, the transmission beam of a BS cannot be perfectly controlled, which potentially causes downlink interference. Thus, coordination between the BS's is necessary to mitigate interference. We propose a cooperative and decentralized interference management scheme that employs spectrum sensing and inter-BS coordination to control the interference in real-time.

To better understand the principles of the CSCN, we present an intuitive scenario in Fig. 6.1, where the CSCN is employed in a metropolitan area to enhance the QoS of users in hotspots. The CSCN BS's provide coverage to the desired area while the interference caused to other areas is controlled. Neighboring BS's coordinate with each other to allocate spectrum resources according to the instantaneous traffic load and reduce interference. From the perspective of both time and space, CSCN improves the spectrum utilization, resulting in the increased network capacity.

Compared with SCN, the spectrum utilization of CSCN is improved due to the adoption of directional antennas and the dynamic resource allocation through BS coordination. As shown in Fig. 6.2, two close BS's with directional antennas, i.e., BS 3 and BS 4, can simultaneously utilize the same spectrum band since there is no overlap between their coverage areas. For BS 1 and BS 2 with overlapped coverage, they could sense the spectrum environment and the instantaneous traffic conditions, and exchange information with each other via backhaul signaling. After that, BS 1 and BS 2 can coordinate with each other to avoid mutual interference and traffic congestion.

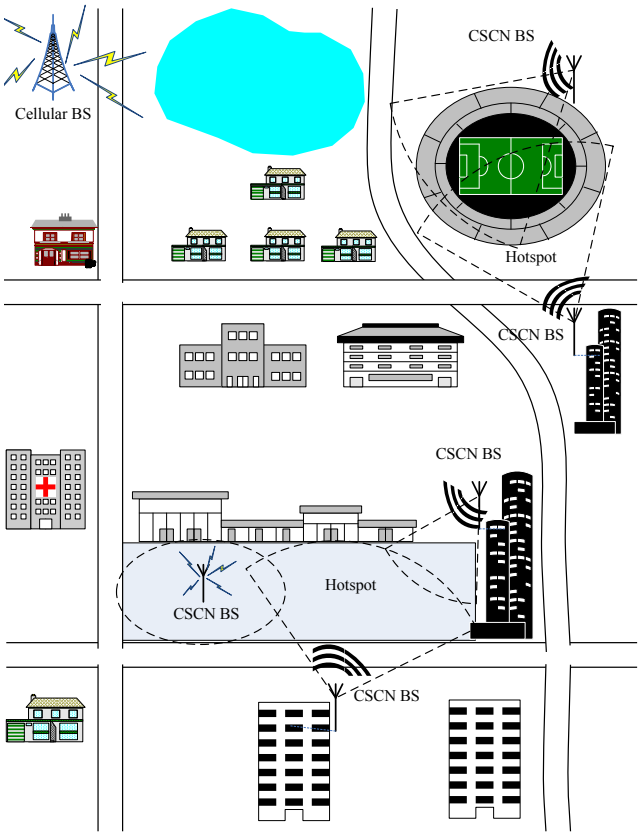

Figure 6.1: An example of CSCN in a metropolitan area.

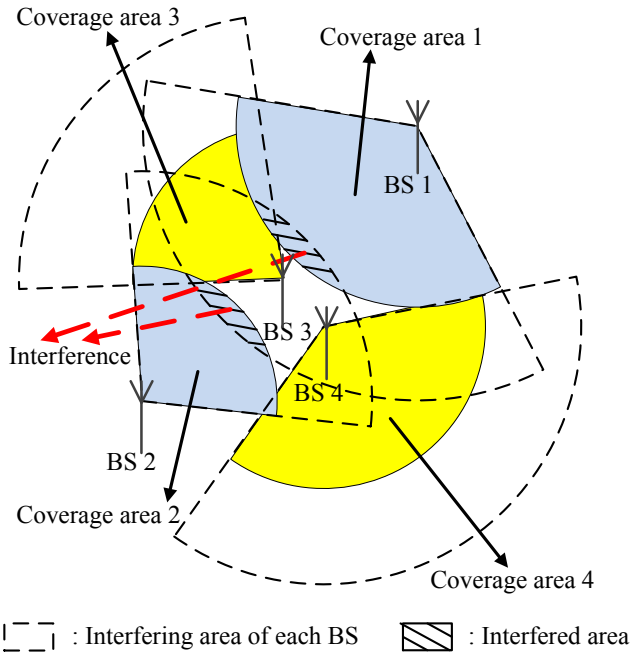

Figure 6.2: Application of directional antenna for BS deployment in the CSCN.

6.4 Technical Aspects

6.4.1 Base Station Deployment

The goal of CSCN BS deployment is to improve spatial reuse as well as mitigating downlink interference through the application of directional antennas. Through careful design on the

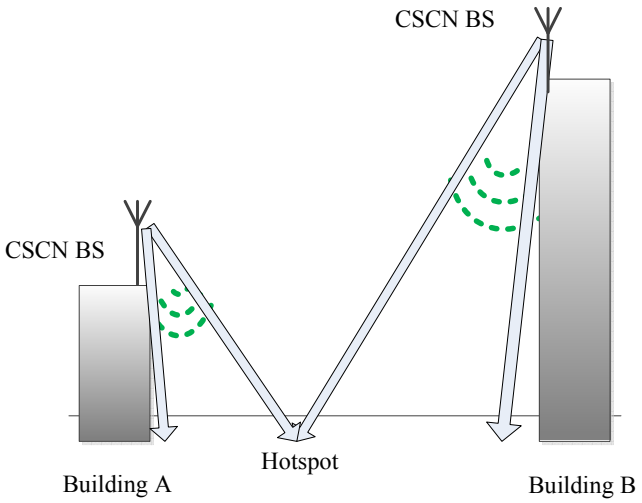

Figure 6.3: Application of tilted antennas to cover a hotspot in the CSCN.

location and coverage area of each BS, the spectrum utilization can be improved to satisfy the high capacity requirement from large number of users.

The deployment of antennas in CSCN can take advantage of the specific architectural layout. Note that, the coverage area of a BS can be controlled by using tilted antenna with adjusted tilt angle [151], as illustrated by the example given in Fig. 6.3. Suppose the zone between the two buildings is a hotspot, and the BS's at the top of buildings A and B provide service to users in the hotspot, with the vertical antenna transmission pattern illustrated in the figure. The specific coverage areas of the two BS's can be preset, or can be distributively coordinated between the two BS's, depending on the technology availability. This way, we can first divide a hotspot into several areas, then employ multiple BS's to serve each area with the above pattern. For BS's that serve the same area, they can coordinate with each other to optimize the network performance and improve the QoS of users in the area.

Next, within one area, we propose a preliminary architecture of BS deployment to mitigate the downlink interference. The proposed architecture bears a decentralized feature that can reduce the control overhead. Take a rectangular area (which may be a square or a lobby in a real-world scenario) shown in Fig. 6.4 as an example. Suppose with a transmission range of d meters, the received signal strength from a CSCN BS is sufficiently small that causes negligible interference. Then, we divide the rectangular area into three kinds of regions with parameter d. To mitigate the interference to the users outside the rectangular area, we add constraints

on antenna directions for BS's located in regions 1 and 2. For a BS in region 1, the antenna azimuth is restricted within a right angle and the radiation direction is limited to the interior zone of the rectangular area, in order to guarantee that the BS does not cause interference to users outside the rectangular area. Similarly, antennas of BS's in region 2 can only direct to the inner side to avoid interference to the outside. For BS's in region 3, due to the relatively longer distance to areas outside the rectangular area, there is no restriction on the directions of antennas. With such constraints, the downlink interference only occurs within users and BS's in the given area, and thus, local coordination among neighboring BS's is feasible.

For an area with a general shape, we first find the reference line that is d meters away from the boundary of the area. Then, according to the reference line, we divide the area into three kinds of regions, namely corner, border, and center. A corner region is between the edges and angles on the boundary and the reference line. A border region is between the edge on the boundary and the reference line. A center region is inside the reference line. These three kinds of regions correspond to regions 1, 2, and 3 in the rectangular example, respectively. Similar to the rectangle area example, we add constraints on the antenna directions of BS's to mitigate the downlink inter-area interference. In the corner region, the antenna azimuth of a BS is restricted by the angle based on the shape of the region, and the radiation direction is limited to the interior zone of the area. In the border region, the antenna of a BS can only be directed to the inner side of the area. In the center region, there is no restriction on the direction of antenna.

The proposed BS deploying architecture not only controls the interference, but also enables spectrum reuse. As the mutual downlink interference between different areas is controlled, each area can utilize all the available downlink spectrum bands, which in turn improves the network capacity.

6.4.2 Dynamic Resource Management

Although the BS deployment could mitigate inter-area interference, the intra-area interference remains to be a problem. Since providing seamless coverage is the prior target, the coverage areas of different CSCN BS's in the same area may overlap, resulting in intra-area interference.

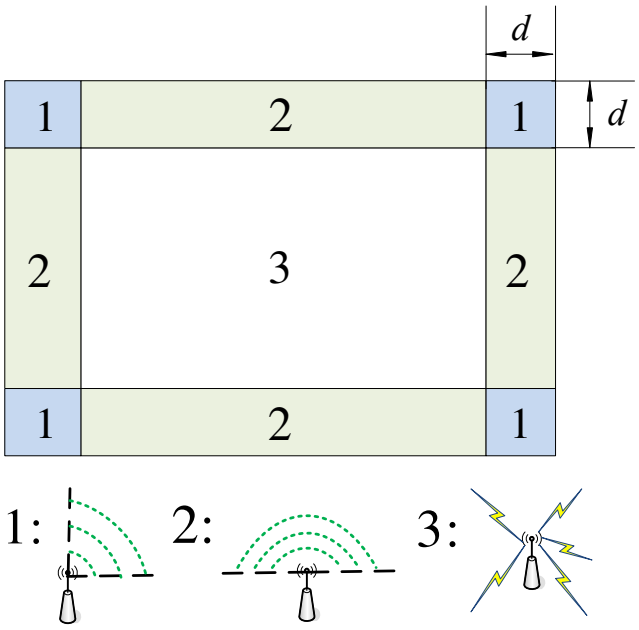

Figure 6.4: Area division for BS deployment in the CSCN.

In this part, we propose spectrum allocation strategies to deal with this problem. We first propose an initial spectrum allocation strategy to mitigate the downlink interference among BS's in the same area. Then, based on the initial spectrum allocation, we propose a cooperative and decentralized BS coordination mechanism to improve the performance of CSCN in the hotspot. When the instantaneous traffic load is changing, CSCN BS's can dynamically adjust the occupied spectrum resource by the antennas through backhaul signaling, and can coordinate with each other to meet the traffic demand.

Initial Spectrum Allocation

In the previous section, we divide a hotspot into several areas and consider the BS deployment pattern for each area. Within such an area, given the coverage areas of all the BS's, we propose a spectrum allocation strategy that considers both spatial reuse and interference mitigation. The objective is to maximize the total number of channels of all BS's, under constraints on interference mitigation and guaranteeing fairness among the BS's. We assume that there are M CSCN BS's and N channels in this area. As shown in Fig. 6.2, each CSCN BS has a coverage area and an interfering area. Define *interference list* as an $M \times M$ binary matrix with element F_{ij}, where $F_{ij} = 1$ indicates that BS i and j interfere with each other due to overlapped

coverage areas, and $F_{ij} = 0$ indicates that BS i and j do not interfere with each other, so they can utilize the same channel.

To avoid the situation that some BS's are allocated with all the channels while some other BS's have no available channel, the number of channels allocated to each BS m should have an upper bound, denoted by A_m. Therefore, the spectrum allocation problem can be formulated as

$$\begin{aligned} \max_{\{\delta_m^n\}} \ & \sum_{n=1}^{N} \sum_{m=1}^{M} \delta_m^n, \\ \text{s.t.} \ & \delta_i^n + \delta_j^n \leq 2 - F_{ij}, \ \forall i, j \in \{1, 2, ..., M\}, \forall n \in \{1, 2, ..., N\} \\ & \sum_{n=1}^{N} \delta_m^n \leq A_m, \ \forall m \in \{1, 2, ..., M\}, \end{aligned} \tag{6.1}$$

where δ_m^n is an indicator and defined as

$$\delta_m^n = \begin{cases} 1, & \text{channel } n \text{ is assigned to BS } m, \\ 0, & \text{otherwise.} \end{cases} \tag{6.2}$$

Obviously, problem (6.1) is a 0–1 integer programming problem with linear constraints. Finding the optimal solution of (6.1) through exhaustive searching incurs NP-hard complexity, since the number of all possible solutions is 2^{MN}. Nevertheless, the solution to the optimization problem is an initial spectrum allocation result that does not consider the traffic load. Therefore, a suboptimal solution is still useful. To obtain a suboptimal solution with low complexity, we relax the constraint $\delta_m^n \in \{0, 1\}$ to $\delta_m^n \in [0, 1]$. Hence, the original problem is transformed into a linear programming problem (LP) and can be solved with existing methods such as the simplex algorithm. The solutions will be rounded up to 1 or down to 0 while satisfying the constraints in (6.1) to obtain a feasible suboptimal solution.

Obviously, A_m directly affects the spectrum allocation result. Note that neighboring BS's cannot simultaneously utilize the same channel, while the objective function of the optimization problem is the total number of channels of all BS's. Therefore, a BS with more neighboring

BS's will be allocated with less channels to leverage spatial reuse and achieve a large object value. With the constraint of A_m, even the BS with most neighboring BS's is expected to be allocated with certain amount of channels. There is a tradeoff between fairness and overall network performance. When A_m is small that approaches $N/2$, the total number of channels allocated to all BS's is small. When A_m is large, the fairness among the BS's will become poor.

The derivation of the channel allocation constraint is based on the downlink scenario, since it is based on the analysis of BS coverage areas. Due to the mobility of users and the omni-directional transmission pattern of uplink signals, the uplink interference cannot be perfectly controlled. To this end, a cooperative interference avoidance scheme will be proposed to address this problem. The goal of adding the channel allocation constraints is to reduce the expected overhead for interference management in the operating process, since neighboring BS's have already been allocated with exclusive channels and the ratio of users that requires interference coordination will be small.

Load Balancing Through Base Station Coordination

In this part, we consider the instantaneous traffic for spectrum allocation. To deal with the potential overloading problem in hotspot, we propose two cooperative and decentralized BS coordination strategies as follows.

- *Channel borrowing*: When a CSCN BS is overloaded, it sends a request for more channels to all the neighboring BS's. Upon receiving the request, the neighboring BS's with idle channels will feedback the idle channel information. Then, the overloaded BS will utilize the channels that all its neighboring BS's reported as idle. These additional channels can assist the overloaded BS to deal with the heavy traffic. With this mechanism, the BS's with heavy traffic could borrow channels from neighboring BS's, and the neighboring BS's utilize the same channel at different time instants.

- *User handover*: Although the channel borrowing mechanism can offer additional channels to an overloaded BS, another problem emerges when the traffic load of a BS keeps increasing. Since a CSCN BS needs to allocate a limited power among all the channels,

with the limited power budget, the power allocated to each channel may not be enough to guarantee the QoS of users. User handover could be an effective solution to this problem. When a CSCN BS detects that the number of users within its coverage area is too high such that a handover is necessary, it sends a request for handover to all the neighboring BS's. Then, the neighboring BS's with available power and spectrum resource will respond to the request. With the feedback information, the overloaded BS reduces the pilot signal strength, part of the users will switch to some neighboring BS's.

6.4.3 Interference Coordination

The proposed deployment and spectrum allocation methods can only deal with the downlink interference. Due to omni-directional uplink transmissions, the uplink interference and interference between the cellular users and the CSCN should also be managed. Furthermore, although the direction antennas can be used to control interference, the potential interference caused by undesired sidelobes should also be mitigated.

We develop a cooperative interference avoidance scheme to mitigate the uplink interference between CSCN BS's. This scheme can also be applied to control the interference between the CSCN and the cellular network with minor modifications. Through spectrum sensing and channel scheduling, interference can be effectively controlled. In particular, the following procedures are adopted by CSCN to achieve uplink interference avoidance.

Spectrum Sensing

We assume that CSCN BS's periodically sense the spectrum occupation of the radio environment. If a CSCN BS (denoted as BS A) detects the uplink signal of a user that is served by another BS, and the signal of this user causes interference to the uplink transmission of BS A, it records the time-frequency usage patterns of this user, and sends the information in a control message to the neighboring BS's. The control message also contains the channel availability information, which indicates the number of remaining channels at the BS.

Confirming the ID of Interfering Users

Once a neighboring BS (denoted as BS B) receives the time-frequency usage patterns from BS A, it compares this information with its uplink scheduling information. If the channel usage pattern coincides with the scheduling information of a user, it indicates that this user interferes with BS A. This way, for each CSCN BS, the nearby interfering users can be identified.

Interference Scheduling

With the channel availability information contained in the control message, BS B compares the numbers of remaining channels of BS A and BS B. If BS A has more remaining channels, BS B informs BS A not to schedule the channel used by the interfering user for uplink transmission to avoid interference. If BS B has more remaining channels, BS B will reschedule the uplink channel for the user that causes interference to BS A.

To avoid the downlink interference caused by BS A to the interfering user, BS B sends the downlink channel scheduling information of the user to BS A. Then, BS A avoids allocating the channel used by this user for downlink transmission.

6.5 Illustrative Example and Simulation Results

To demonstrate the high potential of the proposed approach for capacity improvement and interference mitigation, we present selected simulation results with different schemes applied. Consider a rectangular hotspot with size 200 m $\times$ 400 m, and divide this hotspot into 8 areas with size 100 m $\times$ 100 m. In each area, the number of users is uniformly distributed with mean value μ. The network capacity and outage probability are two criteria for performance evaluation, and we assume that an outage event happens when the capacity of a user falls below a predefined threshold. In this simulation, we focus on the effects of applying deployment pattern and channel borrowing. The downlink transmission is considered since most high data rate services are provided via the downlink. In the simulations, the BS transmission power is 20 dBm and the noise power density is -174 dBm/Hz; the channel bandwidth is 200 kHz and there are 100 channels; the outage threshold is set to 100 kbps; the channel has path loss

$15.3+37.6\log_{10}(R)$ in dB [59] and experiences Rayleigh fading, where R is the transmission distance in meters.

In the first scenario, we consider the original SCN without any modifications. Each area is served by one BS located at the center and the BS adopts an omni-directional antenna. The neighboring BS's use different sets of channels to mitigate interference, and the spectrum usage of each BS is static. In the second scenario, BS's are equipped with directional antennas (DA), and one area is served by multiple BS's with the proposed deployment architecture. For simplicity, we consider each area served by two BS's with directional antennas, a user is served by the BS with the strongest received signal strength. In the third scenario, we enable channel borrowing (CB) between neighboring BS's, and the BS's employ omni-directional antennas. In the last scenario, both DA and CB are employed. We assume that when the number of users is less than the number of channels allocated to the serving BS, each user is allocated with one channel. When the number of users is larger than the number of channels, one channel is shared by multiple users in a TDMA pattern. In detail, suppose the number of users is U, the number of channels is N. When $N < U < 2N$, the BS randomly selects $U - T$ pairs users and assign one channel for each pair of users. When $2N < U < 3N$, the BS randomly select $U - 2N$ groups of users, each group consists of 3 users, and these 3 users share one channel in different time slots. This process can be repeated as the U further increases.

Fig. 6.5 shows the aggregated capacity of all users in the hotspot as a function of different average number of users (μ). The incorporation of both deployment designs with DA and CB between neighboring BS's does improve the capacity compared to the original SCN. The deployment of BS's with directional antennas creates more spatial spectrum reuse opportunities, the number of channels allocated to each user is increased, which in term improves the sum capacity. The CB offers some capacity gain when the average number of users is larger than the number of pre-allocated channels. This is because the borrowed channels from neighboring BS's could assist an overloaded BS to provide better service to users and improve the sum capacity. When the average number of users keeps on increasing, the capacity gain offered by CB becomes limited, since the neighboring BS's are less likely to have any unused channels. When both CB and deployment design with DA are applied, the network capacity is further

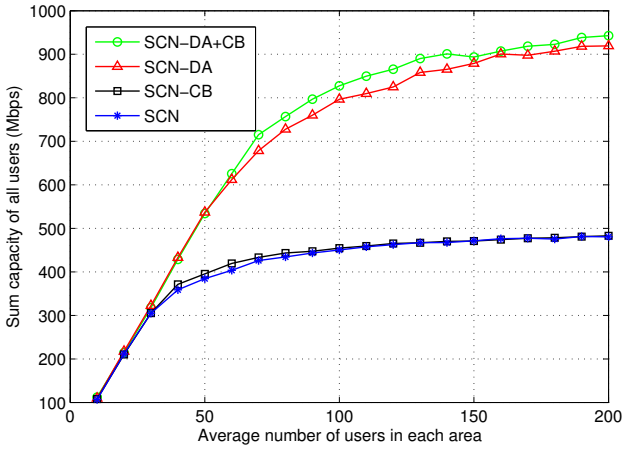

Figure 6.5: Area division for BS deployment in the CSCN.

improved, which results in the further reduced outage probability. Therefore, we conclude that the enhancements adopted in CSCN over SCN can effectively improve the network capacity.

Fig. 6.6 shows the outage probability versus different average number of users (μ). Due to the capability for capacity improvement, the deployment design and CB also contribute to outage reduction. Note that the performance gain of CB decreases when the average number of users is large, since neighboring BS's are also likely to be overloaded and there is no channel to borrow. From the perspectives of space and time, both schemes create spectrum opportunities and improve spectrum utilization, resulting in less traffic congestion. As expected, the combination of the two schemes could achieve the lowest outage probability. Similarly, we can conclude that the CSCN is effective in serving hotspots with the proper design on the technical aspects as discussed.

6.6 Future Research Directions

Based on the CSCN architecture, the following problems are open issues that should be further investigated to fully harvest the high potential of CSCN.

- *Optimal Base Station Deployment*: In the proposed BS deployment scheme, the downlink interference between different areas is controlled through adding constraints on antenna directions. Each area can thus reuse more spectrum resource, leading to improved capacity. However, one open yet challenging problem is how to optimize the placement

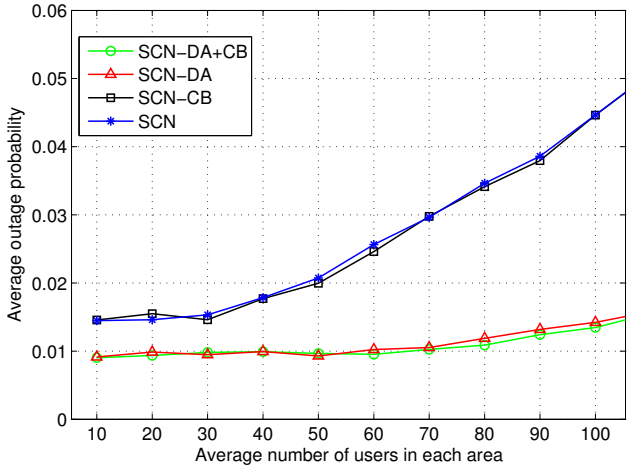

Figure 6.6: Area division for BS deployment in the CSCN.

of BS's, as well as the azimuth and range of antennas, so that the sum capacity can be maximized while the QoS requirements of all users in the hotspot be satisfied. We can employ the cooperation between BS's to optimize the performance, but the technical details require further research.

- *Optimal Handover Strategy*: Due to the relatively small coverage area of a CSCN BS, multiple neighboring BS's can be candidates for load balancing when a CSCN BS is overloaded. The handover decisions determine the capacities of users and interference patterns. Furthermore, the available power and spectrum resources and traffic loads of these neighboring BS's are different. It is thus desirable to develop the optimal handover strategy for best CSCN performance.

- *Power Control*: It is well known that power control is a fundamental approach for capacity improvement and interference mitigation. This chapter does not discuss power control problem due to the lack of space. In future work, cooperative joint power and spectrum allocation algorithms should be developed with easy implementation and low complexity.

In addition, it would also be interesting to develop a CSCN testbed, such that the system performance can be demonstrated under a realistic wireless environment. Such a CSCN testbed can

not only validate the theoretical results, but also reveal new practical constraints that should be considered in the modeling and analysis, as well identifying new research problems.

6.7 Conclusion

In this chapter, we present a CSCN architecture to meet the high capacity demand in hotspots. The proposed architecture adopts sectorized based station deployment, dynamic resource management based on the dynamic traffic load, and interference coordination to achieve enhanced capacity and reduced outage probability. The key is to leverage the cooperation of the neighboring BS's, for better spectrum reuse, interference mitigation, and capacity enhancement. The proposed approach has distributed operation, which is amenable for practical systems. The high potential of the proposed architecture is demonstrated with an illustrative example and simulation results. We conclude this chapter with a discussion of open problems for future research.

Chapter 7

Dealing with Link Blockage with a Combination of Multiple Approaches

7.1 Introduction

With the growing popularity of data-intensive applications, the next generation (i.e., 5G) wireless network is expected to provide 1000x data rate [1]. Millimeter-wave (mmWave) communication is one of the key enabling technologies of 5G wireless to meet such challenges [2, 153], along with massive MIMO [3] and small cells [54, 105]. Operating in the higher end of spectrum ranging from 30GHz to 300GHz, a large bandwidth is available (e.g., a 7 GHz license-free spectrum between 57 GHz and 64 GHz was approved by FCC), resulting in significantly improved data rate.

Despite such great potential, a major challenge of mmWave communications is to overcome blockages. Due to the short wave length, the mmWave cannot penetrate obstacles such as walls and human bodies. Thus, the line-of-sight (LOS) path between a base station (BS) and a user equipment (UE) may be easily blocked due to the mobility of the UE or other UEs. To maintain connectivity, alternative links have to be found and used. To this end, there are three possible approaches, including (i) device to device (D2D) relaying by other UEs, (ii) refection of beams [155, 156], (iii) handover to other BSs. The D2D enabled mmWave network was considered in [158, 160], in which effective MAC protocols were proposed. A multi-beam reflection architecture was recently proposed in [162], where the concept of beamspace MIMO was introduced and a multiplexing gain was achieved. In [163], an inter-BS coordination mechanism is designed to deal with the NLOS challenge by optimizing the set of BSs serving each UE.

Although these solutions are effective in dealing with blockage, the performance gain is limited by some inherent factors of these approaches. For D2D relaying, a relaying UE must share part of its resource with other UEs. Besides, multi-hop relaying increases the delay since the packets have to be forwarded multiple times. The major challenge of multi-beam reflection is the large path loss exponent of NLOS links. Compared to LOS links with a typical path loss exponent of 2, the path loss exponent of NLOS can be as large as 4 [165]. Hence, only a certain set of UEs are suitable for multi-beam reflection. In addition, in case of serving a large number of UEs, the transmission power allocated to each UE would be limited, resulting in degraded quality of service (QoS). The handover approach, which seems easy to implement in traditional cellular network, requires intensive coordination among BSs in an mmWave network. Due to the narrow beam of mmWave transmissions, discovering an alternative BS and tracking roaming UEs may incur additional overhead, such as that caused by beam sweeping [166]. To avoid these drawbacks and fully harness the benefits of these approaches, a combination of multiple approaches with a proper integral design has the potential to enhance the system performance. As the UEs are served by multiple approaches, the number of UEs served by each approach is reduced, the resource allocated to each UE can be increased, resulting in improved performance. Then, how to select the sets of UEs served by different approaches is a key factor that impacts the system performance, but has not been investigated in previous works.

In this chapter, we consider a combination of D2D relaying and multi-beam reflection with an adaptive selection between the two approaches in a single cell time division duplex (TDD) mmWave network. We formulate a joint mode selection and resource sharing problem with the objective of maximizing the sum logarithm rate, and propose a two-stage solution algorithm. In the first stage, we consider the case that all NLOS UEs are served by D2D relaying, and derive the optimal resource sharing solution. Based on the solution, an adaptive mode selection algorithm is then proposed in the second stage to determine the sets of NLOS UEs served by D2D relaying and multi-beam reflection. The proposed scheme is evaluated with simulations and compared with several benchmark schemes, where considerable performance gains are achieved.

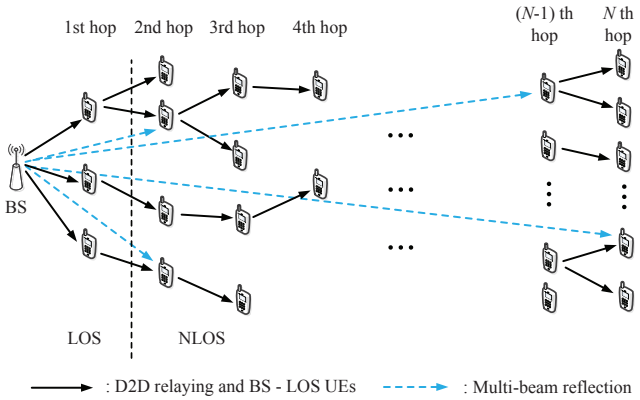

Figure 7.1: System model of a D2D and multi-beam enabled multi-hop mmWave cellular network.

.

In the remainder of this chapter, we present the system model and problem formulation in Section 7.2. The solution algorithm is given in Section 7.3. The simulation results are discussed in Section 7.4. We conclude this chapter in Section 7.5.

7.2 Problem Formulation

We consider an mmWave cellular network that supports directional transmissions with a small beamwidth, e.g., less than 10°. Due to the "pseudo-wired" property and large propagation loss, the performance gain achieved from inter-BS interference coordination is limited. Thus, we focus on link scheduling of *one BS* and the UEs it serves. Assume the set of UEs served by each BS is predetermined, we focus on a tagged BS serving K UEs indexed by $k = 1, 2, \ldots, K$. The UEs are subject to random blockage and can relay data for other UEs in case of blockage. The blocked UEs are served by either multi-hop D2D relaying or multi-beam reflection.

We assume the system operate in the TDD mode, so that channel reciprocity can be exploited for efficient channel estimation. As shown in Fig. 7.1, we consider link scheduling for downlink transmissions. Each NLOS UE receives its packets from a single relaying UE. Then, the network architecture of multi-hop D2D relaying can be viewed as a tree, while multi-beam reflection can be viewed as an alternative (or, augment) connection from the root to a node that corresponds to an NLOS UE.

Define x_k and y_k as indicators for multi-hop D2D relaying and multi-beam reflection for user k, respectively, given as

$$x_k \doteq \begin{cases} 1, \text{ user } k \text{ is served by D2D relaying or LOS link} \\ 0, \text{ otherwise,} \end{cases} \tag{7.1}$$

$$y_k \doteq \begin{cases} 1, & \text{user } k \text{ is served by multi-beam reflection} \\ 0, & \text{otherwise,} \end{cases} \tag{7.2}$$

We assign $x_k = 1$ to LOS UEs since they are part of the D2D relaying architecture. We then use a *descendent matrix* to indicate the relaying route of each NLOS UEs, defined as

$$a_{k,k'} \doteq \begin{cases} 1, & \text{user } k' \text{ is a descendent of user } k \\ 0, & \text{otherwise,} \end{cases} \tag{7.3}$$

Since $a_{k,k'}$ is defined for UEs involved in the D2D relaying process, we have $a_{k,k'} \leq x_k$ and $a_{k,k'} \leq x_{k'}$. Let ρ_k be the depth of UE k in the D2D relaying tree, given by

$$\rho_k = \sum\nolimits_{k' \neq k} a_{k',k} + 1. \tag{7.4}$$

That is, ρ_k is also the number of hops required to transmit the packets of UE k. Let N be the maximum value of ρ_k. To accommodate N-hop D2D relaying and guarantee that the packets of UEs at the Nth hop can be delivered at the end of each downlink transmission, we divide the downlink transmission period into N time slots as shown in Fig. 7.2. Each time slot is used for transmission from UEs in one hop to UEs in the next hop. For a UE served with multi-beam reflection, the entire downlink period is allocated to the UE.

We assume that all UEs can only perform half duplex relaying; two adjacent links that share the same UE cannot operate concurrently [167]. Under this constraint, a total number of $\lceil \frac{N}{2} \rceil$ transmissions can be implemented, as shown in Fig. 7.2. For example, after the 2nd stage of the 1st transmission, which is from UEs with $\rho_k = 1$ to UEs with $\rho_k = 2$, the 2nd

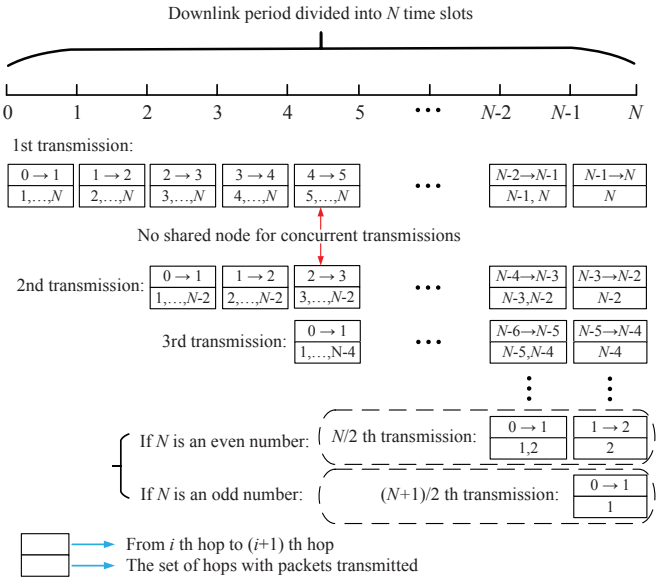

Figure 7.2: Transmission pattern of a TDD-based multi-hop D2D mmWave cellular network.

transmission can be initiated since UEs with $\rho_k = 1$ have finished transmission and are free for reception. Note that, as the index of transmissions increases, the packets of UEs with larger ρ_k are not contained since there is no enough time for the transmissions of these UEs. In particular, the packets of UEs with $\rho_k = N - 1$ and $\rho_k = N$ are not contained in the 2nd transmission, the packets of UEs with $\rho_k = N - 3, ..., N$ are not contained in the 3rd transmission, and so on. Then, the total number of transmissions for UE k, θ_k, is given by

$$\theta_k = \begin{cases} \frac{N}{2} - \left\lceil \frac{\rho_k}{2} \right\rceil + 1, & N \text{ is even} \\ \frac{N+1}{2} - \left\lfloor \frac{\rho_k}{2} \right\rfloor, & N \text{ is odd,} \end{cases} \tag{7.5}$$

The BS is able to serve multiple users on the same time-frequency resource block with advanced techniques, such as hybrid beamforming [154]. Due to hardware constraints, when a UE relays the packets of multiple UEs, resource sharing is required among these UEs. Without loss of generality, we assume time division multiple access (TDMA) is applied when a UE forwards the packets to multiple UEs. Specifically, a fraction of time in each time slot is allocated to each D2D link between the relaying UE and a UE at the next hop. We define $t_k^{i,j}$ as the fraction of time allocated to UE k on its ith hop during the jth transmission. From the

perspective of outflow, $t_k^{i,j}$ should satisfy

$$\sum_{k' \neq k} a_{k,k'} t_{k'}^{\rho_k+1,j} \leq 1, k \in \{k | \rho_k \leq N-1\}, j=1, ..., \theta_k + \left\lceil \frac{\rho_k}{2} \right\rceil . \tag{7.6}$$

For UEs with $\rho_k = 1$, i.e., the LOS UEs, the inflow constraint is given as

$$\sum_{k' \neq k} a_{k,k'} t_{k'}^{1,j} + t_k^{1,j} \leq 1, k \in \{k \,|\rho_k = 1\}, j = 1, ..., \left\lceil \frac{N}{2} \right\rceil . \tag{7.7}$$

The inflow constraint (7.7) is only required for the 1st hop transmission, since the inflow of other hops would only be part of a time slot due to TDMA at the previous hop; the left-hand side of (7.7) would always be no larger than 1. In a multi-hop transmission, the data rate of a link at the current hop should be no less than the data rate of the link at the next hop. Thus, $\{t_k^{i,j}\}$ should also satisfy

$$C_k^i t_k^{i,j} \geq C_k^{i+1} t_k^{i+1,j}, k \in \{k \,|\rho_k \geq 2\}, i = 1, ..., \rho_k - 1, \tag{7.8}$$

where C_k^i is the link capacity of hop i of UE k, given by

$$C_k^i = B \log \left(1 + \gamma_k^i\right), \tag{7.9}$$

where B is the system bandwidth and γ_k^i is the SINR of hop i of UE k. Here, the SINR is constant over all θ_k transmissions, since the duration of uplink and downlink periods is less than the coherence interval in a TDD system.

Next, we derive the SINR expressions under different transmission schemes, employing the baseline SINR model presented in [165]. For UEs served by D2D relaying, we have

$$\gamma_k^i = \frac{p_k^i h_k^i G_k^i (d_k^i)^{-2}}{\sigma^2}, \tag{7.10}$$

where h_k^i is the small scale fading, which is a normalized Gamma random variable [165]; G_k^i is antenna array gain; d_k^i is the distance of the link; and p_k^i is the transmission power of the BS or

UE for the ith hop of UE k. The noise power is σ^2, and we neglect the impact of interference. In (7.10), we assume that the LOS path loss exponent is 2.

For UEs served by BS with multi-beam reflection and UEs served by BS with an LOS link, we assume the power of BS is equally allocated to these UEs. Then, the transmission power allocated to each UE is

$$\tilde{p}_k = \frac{P_0}{\sum_k y_k + N_{\text{LOS}}}, \tag{7.11}$$

where P_0 is the BS power and N_{LOS} is the number of LOS UEs. Similar to (7.10), the SINR of an LOS UE is given by

$$\gamma_k^1 = \frac{\tilde{p}_k h_k^1 G_k^1 (d_k^1)^{-2}}{\sigma^2}. \tag{7.12}$$

Suppose UE k is served with the reflection of M_k beams indexed by $m = 1, ..., M_k$. Let $\tilde{\gamma}_k^m$ be the SINR of the path corresponding to the m th beam of UE k, given as

$$\tilde{\gamma}_k^m = \frac{\tilde{p}_k^m h_k^m G_k^m (d_k^m)^{-4}}{\sigma^2}, \tag{7.13}$$

where $\tilde{p}_k^m$, h_k^m, G_k^m, and d_k^m are the power, small-scale fading, antenna array gain, and distance of the mth path, respectively. We neglect the interference caused by side lobes. Without loss of generality, we assume $\tilde{p}_k$ is equally allocated to the M_k beams. The path loss factor for NLOS reflection is 4.

For UE k served by multi-hop D2D relaying, its downlink data rate is the sum of all θ_k transmissions. In each multi-hop transmission, the effective data rate is determined by data rate of the *final hop* divided by the number of hops. Then, the data rate of UE k when served by multi-hop D2D relaying in the downlink period is given by

$$R_k = \sum_{j=1}^{\theta_k} \frac{C_k^{\rho_k} t_k^{\rho_k, j}}{\rho_k}. \tag{7.14}$$

The data rate of LOS UEs is a special case of (7.14) with $\rho_k = 1$.

When UE k is served by multi-beam reflection in the downlink period, its data rate is given by

$$\tilde{R}_k = \sum_{m=1}^{M_k} B \log\left(1+\tilde{\gamma}_k^m\right). \tag{7.15}$$

In this chapter, we aim to maximize the logarithm rate sum of all UEs with adaptive selection between D2D relaying and multi-beam reflection for all NLOS UEs. Let $\mathbf{x}$, $\mathbf{y}$, $\mathbf{a}$, and $\mathbf{t}$ be the vector/matrix forms of $\{x_k\}$, $\{y_k\}$, $\{a_{k,k'}\}$, and $\{t_k^{i,j}\}$. The problem is formulated as

$$\textbf{P1:} \max_{\{\mathbf{x},\mathbf{y},\mathbf{a},\mathbf{t}\}} \left\{ \sum_{k=1}^{K} x_k \log R_k + \sum_{k=1}^{K} y_k \log \tilde{R}_k \right\} \tag{7.16}$$

subject to:

$$x_k + y_k \leq 1,\ \forall k \tag{7.17}$$

$$\sum_{k'\neq k} a_{k,k'} t_{k'}^{\rho_k+1,j} \leq 1, k \in \{k|\rho_k \leq N-1\}, j=1,...,\theta_k+\left\lceil \frac{\rho_k}{2} \right\rceil \tag{7.18}$$

$$\sum_{k'\neq k} a_{k,k'} t_{k'}^{1,j} + t_k^{1,j} \leq 1, k \in \{k\,|\rho_k = 1\}, j=1,...,\left\lceil \frac{N}{2} \right\rceil \tag{7.19}$$

$$C_k^{i-1} t_k^{i-1,j} \geq C_k^i t_k^{i,j},\ k \in \{k\,|\rho_k \geq 2\}, i=1,...,\rho_k \tag{7.20}$$

$$\sum\nolimits_{k'\neq k} a_{k',k} \leq 1,\ k \in \{k\,|\rho_k \leq 2\}, \tag{7.21}$$

$$a_{k,k'} \leq x_k,\ \forall k. \tag{7.22}$$

$$t_k^{i,j} \leq x_k,\ 0 \leq t_k^{i,j} \leq 1,\ \forall k,i,j. \tag{7.23}$$

$$x_k, y_k, a_{k,k'} \in \{0,1\},\ \forall k. \tag{7.24}$$

Constraint (7.20) is due to the fact that each node in the D2D relaying tree can only have one parent node.

7.3 Solution Algorithm

Problem **P1** is a mixed integer programming problem with multiple sets of variables, which cannot be solved with standard techniques. To derive an effective solution, we propose a two-stage algorithm for adaptive selection between D2D relaying and multi-beam reflection. In the

first stage, we consider the case that all NLOS UEs are served with D2D relaying and derive the optimal resource allocation solution. Based on the solution, the UEs that can be served by multi-beam reflection are evaluated in the second stage, then the set of UEs to be served by multi-beam is determined.

7.3.1 First Stage

We assume that path selection for D2D relaying is pre-determined with a routing approach, e.g., adopting Dijkstra's Algorithm by setting the weight of each link as the inverse of its channel gain [168]. With all NLOS UEs served by D2D relaying, problem **P1** is reduced to the following resource allocation problem.

$$\textbf{P2:} \max_{\{\mathbf{t}\}} \sum_{k=1}^{K} \log \left(\sum_{j=1}^{\theta_k} \frac{C_k^{\rho_k} t_k^{\rho_k, j}}{\rho_k} \right) \tag{7.25}$$

$$\text{subject to: } (7.18) - (7.20)$$

It can be verified that problem **P2** is a convex optimization problem and strong duality holds. The Lagrangian dual method can be applied to obtain the optimal solution.

7.3.2 Second Stage

Based on the optimal solution of problem **P2**, we then evaluate the performance gain of switching a UE from D2D to multi-beam, and determine the set of UEs to be served with multi-beam reflection. We define $\Delta_k^{[\tau]}$ as the performance gain of all UEs by selecting UE k to switch from D2D relaying to multi-beam reflection at the τth round. For a UE with descendent UE(s) in the D2D tree, we assume that all of its descendent UEs would also switch to multi-beam reflection since re-constructing the D2D tree is time-consuming, which should not be carried out frequently. Then, the performance gain is given as: $\Delta_k^{[\tau]} = \log \tilde{R}_k^{[\tau]} - \log R_k^{[\tau]} + \sum_{k' \neq k} a_{k,k'} \left(\log \tilde{R}_{k'}^{[\tau]} - \log R_{k'}^{[\tau]} \right)$.

Algorithm 11: Adaptive Mode Selection Algorithm

Initialize: $x_k = 1, y_k = 0, \forall k$, $\Phi = \{1, ..., K\}$;
$\tau = 1$, $l^{[1]} = \arg\max_{\{k|x_k=1\}} \Delta_k^{[1]}$;
while $\Delta_l^{[\tau]} > 0$ **do**
  if $\rho_l = 1$ **then**
    Set $y_l = 1$, $x_l = 0$, $\Phi = \Phi - \{l\}$;
    for $k = 1 : K$ **do**
      Update $R_{k'}^{[\tau]}$ with (7.14) by adding $t_l^{\rho_l,j} / \sum_k x_k$ to each $t_k^{\rho_k,j}$, $k \in \Phi$;
      Update $\tilde{R}_k^{[\tau]}$ with (7.15) by multiplying $\tilde{p}_k$ by $(\sum_k y_k + N_{\text{LOS}})/(\sum_k y_k + 1 + N_{\text{LOS}})$;
      Calculate $\Delta_k^{[\tau]}$;
    end
  else
    Set $y_l = 1$, $x_l = 0$; $y'_l = 1$, $x'_l = 0$, for $l' \in \eta_l$;
    $\Phi = \Phi - \{l\} - \eta_l$;
    for $k = 1 : K$ **do**
      Update $R_{k'}^{[\tau]}$ with (7.14) by adding $(t_l^{\rho_l,j} + \sum_{l' \neq l} a_{l,l'} t_{l'}^{\rho_{l'},j}) / \sum_k x_k$ to each $t_k^{\rho_k,j}$, $k \in \Phi$;
      Update $\tilde{R}_k^{[\tau]}$ with (7.15) by multiplying $(\sum_k y_k + N_{\text{LOS}})/(\sum_k y_k + 1 + \sum_{l' \neq l} a_{l,l'} + N_{\text{LOS}})$ to each $\tilde{p}_k^m$, $k \notin \Phi$;
      Calculate $\Delta_k^{[\tau]}$;
    end
  end
  $l^{[\tau]} = \arg\max_{\{k \in \Phi\}} \Delta_k^{[\tau]}$;
  $\tau = \tau + 1$;
end

Since the objective function is to maximize the sum logarithm rate, the value of $t_k^{\rho_k,j}$ would be close to each other among different UEs. Thus, when an NLOS UE l switches from D2D relaying to multi-beam reflection, we approximate the values of $t_k^{\rho_k,j}$ of other UEs to be increased by $\frac{t_l^{\rho_l,j}}{\sum_k x_k}$ or $\frac{t_l^{\rho_l,j} + \sum_{l' \neq l} a_{l,l'} t_{l'}^{\rho_{l'},j}}{\sum_k x_k}$, depending on whether UE l has any descendent in the D2D tree. Since we assume equal power allocation, the transmission power allocated to each LOS UE and multi-beam NLOS UE is decreased by a factor of $\frac{\sum_k y_k + N_{\text{LOS}}}{\sum_k y_k + 1 + N_{\text{LOS}}}$ or $\frac{\sum_k y_k + N_{\text{LOS}}}{\sum_k y_k + 1 + \sum_{l' \neq l} a_{l,l'} + N_{\text{LOS}}}$, depending on whether UE l has any descendent in the D2D tree.

Let $\eta_k = \{k' | \, a_{k,k'} = 1\}$ and Φ be the set of UEs served by D2D relaying. The adaptive algorithm for selection between D2D and multi-beam is summarized in Algorithm 11. In each round of Algorithm 11, the UE or the set of UEs that brings the largest performance gain is selected to switch from D2D relaying to multi-beam reflection. Such process terminates until no positive gain can be achieved.

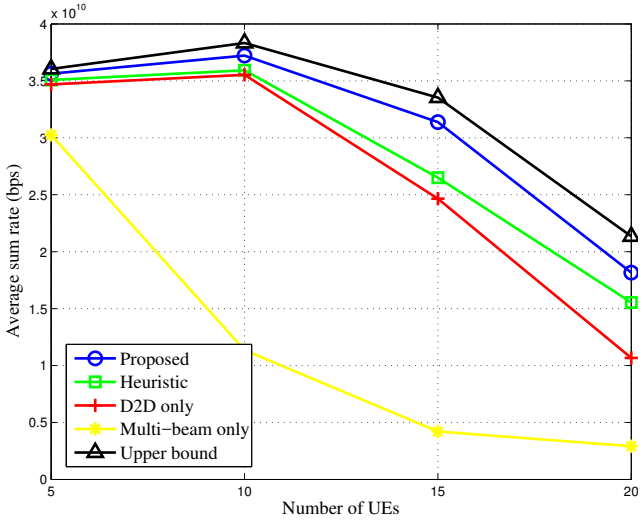

Figure 7.3: Average sum rate under different numbers of UEs. $\kappa = 0.02$.

7.4 Simulation Study

We validate the performance of the proposed scheme through MATLAB simulations. We consider one BS serving multiple UEs with coverage range of 50 m. The system bandwidth is 1 GHz. The UEs are randomly distributed in the coverage area. Each UE is subject to random blockage with probability P_k^b. We assume that the blockage probability is proportional to the distance between the UE and the BS with coefficient κ, given as $P_k^b = \min\{\kappa \cdot D_k, 1\}$. As an example, when $\kappa = 0.04$, a UE that is 10 m away from the BS has a probability of 0.4 to be blocked; for UEs that are more than 25 m away, they are always blocked. We compare with a heuristic scheme. In each round of the heuristic scheme, the UE with the largest value of $\tilde{R}_k$ is selected to be served with multi-beam reflection. Such process terminates until $\sum_k y_k \tilde{R}_k$ decreases. We also consider the case of setting the objective function of Problem **P2** as sum rate maximization, which serves as an *upper bound* for the sum rate performance.

Fig. 7.3 shows the sum rate performance under different number of UEs. The sum rate of the multi-beam only scheme is lower than other schemes due to the large path loss exponent of NLOS links and the increased transmission distance brought by reflection. The D2D only scheme achieves higher sum rate than the multi-beam only scheme, but its performance is significantly degraded when the number of UEs is large. The degraded performance of D2D only scheme results from both resource sharing among UEs and the increased time spent on

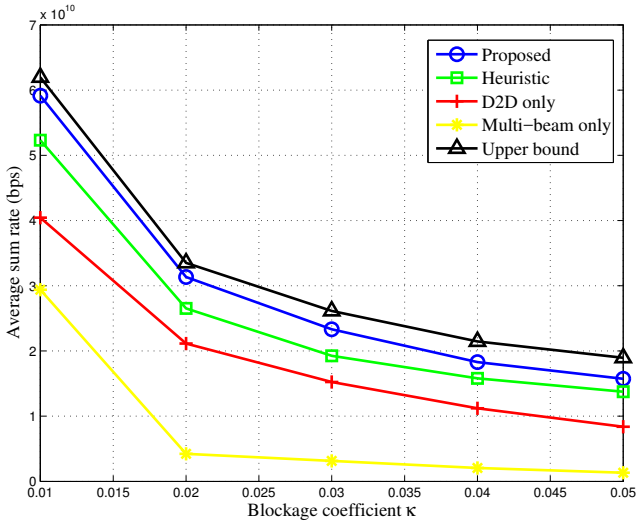

Figure 7.4: Average sum rate under different values of κ. The number of UEs is 15.

multi-hop transmission. By selecting some UEs to be served by multi-beam reflection, the heuristic scheme and the proposed scheme achieve better performance, due to the reduced number of UEs involved in D2D transmission. The proposed scheme outperforms the heuristic scheme since the proposed adaptive algorithm jointly considers the performances of UEs served by D2D and multi-beam, while the heuristic scheme is only based on UEs served by multi-beam. The performance of the proposed scheme is close to its upper bound, showing that the performance loss due to fairness concern is relatively small. The results of Fig. 7.3 indicate that both D2D relaying and multi-beam reflection are highly limited by the increasing traffic. A combination of the two approaches with proper UE selection can effectively improve the system performance.

The performances under different blockage coefficients, κ, is shown in Fig. 7.4. As κ increases, the performances of all schemes are degraded due to increased ratio of NLOS UEs. It can be seen that the performance of both D2D only and multi-beam only schemes are highly sensitive to blockage, since the time and power resources are shared by the increasing number of NLOS UEs. With adaptive selection between D2D and multi-beam, the proposed scheme achieves considerable performance gain, especially when κ is large.

Fig. 7.5 shows the fairness performance by setting different objective functions for Problem **P2**. We use the Jain's fairness index, $\left(\sum_k x_k R_k + \sum_k y_k \tilde{R}_k\right)^2 / K \cdot \sum_k (x_k R_k + y_k \tilde{R}_k)^2$, to measure fairness between different UEs. It can be seen from Fig. 7.5 that when the objective is

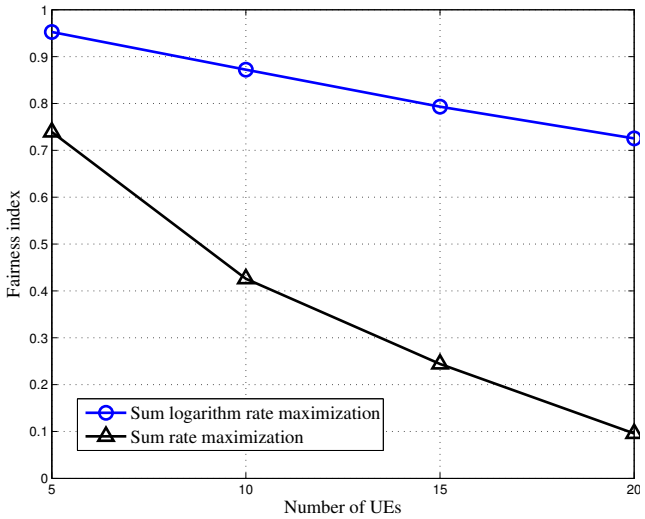

Figure 7.5: Fairness with different objective functions. The number of UEs is 15 and $\kappa = 0.02$.

set to sum rate maximization, the fairness is poor, since the links with high channel gain would be allocated with much more transmission time than the links with low channel gain. With the objective of sum logarithm rate maximization, the proposed scheme achieves a good tradeoff between system performance and fairness.

7.5 Conclusions

We considered a combination of D2D relaying and multi-beam reflection to overcome blockage as well as enhance performance of a TDD mmWave network. A two-stage solution algorithm was proposed to determine the set of UEs served by D2D relaying and multi-beam reflection. The effectiveness of the proposed scheme was validated with simulations.

Chapter 8

Conclusions

In this dissertation work, we propose efficient solutions for several key technologies of 5G wireless networks. We aim to enhance the system performance with particular focus on dealing with number of users. In addition to each individual technology, we also investigate how to enable efficient integration of different technologies.

In Chapter 2, we investigate the problem of BS sleep control in massive MIMO HetNets. An integer programming problem formulation is given, followed by a centralized solution and two distributed schemes. Though our analysis and simulation results, we find that the system energy efficiency can be significantly improved while the data rate loss due to dynamic BS sleep control is relatively small.

In Chapter 3, we consider interference management of massive MIMO HetNets from the perspective of antenna array processing. We employ a non-uniform antenna placement configuration called nested array, which can achieve $O(N^2)$ DoF with N antennas. Then, the design issue is how to use the DoF of each BS for service provision and interference nulling. We formulate an integer programming and propose approximation based solution. The results show that the optimization of DoF use can effectively enhance the data rate performance.

In Chapter 4, we consider joint frame design, resource allocation, and user association to optimize the performance of a massive MIMO HetNet with wireless backhaul. We formulate the problem as an integer programming and propose an iterative solution algorithm. With adaptive pilot length, i.e., the number of symbols dedicated to pilots in each frame, the proposed scheme outperforms other schemes with fixed frame structures.

In Chapter 5, we apply full-duplex transmission into a femtocell network and propose a joint duplex mode selection, channel allocation, and power control scheme to enhance the system sum rate. We employ a stable roommate matching algorithm to determine the pairing strategy of users and make a selection between HD and FD based on the pairing result. With the proposed user pairing, adaptive HD/FD selection, and resource allocation, the interference caused by FD transmission is effectively controlled and the system performance is improved compared to the schemes without user pairing and adaptive mode selection.

In Chapter 6, we present a cooperative small cell network architecture to mitigate interference and improve the capacity of a dense small cell network. The key components include adaptive BS deployment and configuration, dynamic resource allocation, and interference coordination. With efficient spectrum reuse and traffic-aware scheduling, the proposed architecture effectively improves the data rate performance of small cell network when serving large number of users.

In Chapter 7, we consider a combination of multiple approaches for dealing with link blockage, with the objective to enhance the performance of a mmWave system. We design an adaptive scheme to select the set of users served by D2D relaying and multi-beam reflection. Simulation results demonstrated that a combination of multiple approaches provides significant performance gain with the proposed scheme.

Chapter 9

Future work

The development of wireless networks is triggered by the emergence of applications that require efficient wireless transmission. While my past research focuses on cellular systems, I plan to explore the wireless networking problems in recently emerged application scenarios, e.g, IoT, cloud/fog computing, etc.

From the perspective of problem-solving techniques, machine learning opens new opportunities to the optimization of wireless networks. While traditional frameworks rely on the availability of network information, there is always a tradeoff between overhead and performance. With machine learning-based approaches, the methodology of network optimization is transformed from planning to learning, which offers multi-fold benefits. For example, the network schedule with reinforcement learning is based on a trial and error process, which does not require explicit or instantaneous network information. On the other hand, the entire network features and the inter-dependent patterns of different nodes can be learned with the trial and error process by multiple times. Thus, compared to greedy and distributed algorithms that are based on limited network information, a reinforcement learning based algorithm have the potential to provide better performance. Due to such promising prospect, I plan to apply machine learning to the design of wireless networks in future research.

The detailed future research directions are as follow.

- **Machine Learning Aided MmWave Network.** The mmWave transmissions is characterized by high data rate and unreliable connection. In contrast, the backhaul link provides relatively low data rate transmission but reliable connection. Then, how to

efficiently assign the capacity of backhaul link to multiple mmWave links is a challenging task. If the transmission content and blocking pattern of each mmWave link can be learned, the scheduling of backhaul link can be optimized to enhance the QoS of mmWave users. I plan to use a deep reinforcement learning approach to tackle this problem and achieve the potential performance gain brought by learning algorithms.

- **Massive IoT Device Scheduling.** The future ubiquitous deployment of IoT devices generates a huge amount of data and a majority of the data would be transmitted via wireless links. For some applications with stringent delay requirements, e.g., emergency report, health monitoring, it is necessary to minimize the delay and guarantee the timeliness of information. Based on my experience on dealing with large number of users/devices, I will work on the transmission strategies and protocols to optimize the delay performance in IoT devices-based wireless networks. As machine to machine (M2M) communication is likely to be applied between IoT devices, such a network architecture brings new challenges and I shall work on related issues as well.

- **Efficient Communication in Fog Computing Based Network.** Due to the expected large volume of data in 5G network, uploading all the data to the cloud is time-consuming and thus not cost-effective. Fog computing, or mobile edge computing, is a promising approach to improve computing efficiency by using local computing units for traffic offloading. A key performance measure for fog computing is the average delay experienced by users, which not only depends on the computing capability, but also depends on efficient communication links that support fast data transmission. Therefore, the scheduling issues such as task assignment and resource allocation should be considered from multiple aspects such as link quality, traffic load, and computing capability. As multiple fog nodes coexist with a cloud node, such an architecture is similar to a two-tier heterogeneous network with multiple small cells. How to efficiently use both kinds of nodes would also be part of my future research.

- **QoE Aware IoT and 5G.** Quality of experience (QoE), which indicates the level of satisfaction of a user, has become a new performance metric in network design. The QoE can

be determined by a combination of multiple QoS parameters, such as outage probability, delay, downloading rate, video type and quality, and energy consumption of mobile devices. The evaluations regarding all these parameters vary over different user. Apart from these parameters, the QoE can also be estimated by evaluating the emotion-aware response from users, which can be collected from IoT devices. As QoE opens a new paradigm of wireless network design, I plan to investigate QoE improvement approaches in 5G and IoT applications. The integration of emotion-aware IoT applications and QoE would be an interesting topic to be studied as well.

References

[1] Qualcomm, "The 1000x data challenge," [online] Available: https://www.qualcomm.com/1000x.

[2] J. Andrews, S. Buzzi, W. Choi, S. V. Hanly, A. Lozano, A. C. K. Soong, and J. C. Zhang, "What will 5G be?" *IEEE J. Sel. Areas Commun.,* vol.32, no.6, pp.1065–1082, June 2014.

[3] M. Feng and S. Mao, "Harvest the potential of massive MIMO with multi-layer techniques," *IEEE Network,* vol.30, no.5, pp.40–45, Sept./Oct. 2016.

[4] M. Feng, S. Mao, and T. Jiang, "BOOST: Base station on-off switching strategy for energy efficient massive MIMO HetNets," in *Proc. IEEE INFOCOM'16,* San Francisco, CA, Apr. 2016, pp.1395–1403.

[5] M. Feng, S. Mao, and T. Jiang, "BOOST: Base station on-off switching strategy for green massive MIMO HetNets," *IEEE Trans. Wireless Commun.,* vol.16, no.11, pp.7319–7332, Nov. 2017.

[6] A. Adhikary, H.S. Dhillon, and G. Caire, "Massive-MIMO meets HetNet: Interference coordination through spatial blanking," *IEEE J. Select. Areas Commun.,* vol.33, no.6, pp.1171–1186, June 2015.

[7] K. Zheng, L. Zhao, J. Mei, B. Shao, W. Xiang, and L. Hanzo, "Survey of large-scale MIMO systems," *IEEE Commun. Sur. & Tut.*, vol.17, no.3, pp.1738–1760, Third Quarter 2015.

[8] T.L. Marzetta, "Noncooperative cellular wireless with unlimited numbers of base station antennas," *IEEE Trans. Wireless Commun.,* vol.9, no.11, pp.3590–3600, Nov. 2010.

[9] H.Q. Ngo, E.G. Larsson, and T.L. Marzetta, "Energy and spectral efficiency of very large multiuser MIMO systems," *IEEE Trans. Commun.,* vol.61, no.4, pp.1436–1449, Apr. 2013.

[10] Y. Xu, G. Yue, and S. Mao, "User grouping for massive MIMO in FDD systems: New design methods and analysis," *IEEE Access Journal*, vol.2, no.1, pp.947–959, Sept. 2014.

[11] E. Björnson, M. Kountouris, and M. Debbah, "Massive MIMO and small cells: Improving energy efficiency by optimal soft-cell coordination," in *Proc. ICT'13,* Casablanca, Morocco, May 2013, pp.1–5.

[12] D. Bethanabhotla, O.Y. Bursalioglu, H.C. Papadopoulos, and G. Caire, "User association and load balancing for cellular massive MIMO," in *Proc. IEEE Inf. Theory Appl. Workshop 2014,* San Diego, CA, Feb. 2014, pp.1–10.

[13] Y. Xu and S. Mao, "User Association in Massive MIMO HetNets," *IEEE Systems Journal*, vol.11, no.1, pp.7–19, Mar. 2017.

[14] D. Liu, et al., "Distributed energy efficient fair user association in massive MIMO enabled HetNets," *IEEE Comm. Lett.,* vol.19, no.10, pp.1770–1773, Oct. 2015.

[15] M. Feng and S. Mao, "Interference management and user association for nested array-based massive MIMO HetNets," *IEEE Trans. Veh. Technol.,* vol.67, no.1, pp.454–466, Jan. 2018.

[16] M. Feng and S. Mao, "Adaptive pilot design for massive MIMO HetNets with wireless backhaul," in *Proc. IEEE SECON'17,* San Diego, CA, June 2017, pp.1–9.

[17] Y. Chen, S. Zhang, S. Xu, and G.Y. Li, "Fundamental tradeoffs on green wireless networks," *IEEE Commun.,* vol.49, no.6, pp. 30–37, June 2011.

[18] E. Oh, B. Krishnamachari, X. Liu, and Z. Niu, "Toward dynamic energy-efficient operation of cellular network infrastructure," *IEEE Commun.,* vol.49, no.6, pp.56–61, June 2011.

[19] M. Feng, S. Mao, and T. Jiang, "Base station ON-OFF switching in 5G wireless networks: Approaches and challenges," *IEEE Wireless Commun.,* vol.24, no.4, pp.46–54, Aug. 2017.

[20] V. Chandrasekhar and J.G. Andrews, "Spectrum allocation in tiered cellular networks," *IEEE Trans. Commun.,* vol.57, no.10, pp.3059–3068, Oct. 2009.

[21] S. Zhang, J. Gong, S. Zhou, and Z. Niu, "How many small cells can be turned off via vertical offloading under a separation architecture?" *IEEE Trans. Wireless Commun.,* vol.14, no.10, pp.5440–5453, Oct. 2015.

[22] S. Cai, Y. Che, L. Duan, J. Wang, S. Zhou, and R. Zhang, "Green 5G heterogeneous networks through dynamic small-cell operation," *IEEE J. Sel. Areas Commun.,* vol.34, no.5, pp.1103–1115, May 2016.

[23] E. Bjö rnson, E.G. Larsson, and M. Debbah, "Massive MIMO for maximal spectral effciency: How many users and pilots should be allocated?" *IEEE Trans. Wireless Commun.,* vol.15, no.2, pp.1293–1308, Feb. 2016.

[24] E. Bjö rnson, L. Sanguinetti, and M. Kountouris, "Deploying dense networks for maximal energy efficiency: Small cells meet massive MIMO," *IEEE J. Select. Areas Commun.,* vol.34, no.4, pp.832–847, Apr. 2016.

[25] H. Ngo, H. Suraweera, M. Matthaiou, and E. Larsson, "Multipair full-duplex relaying with massive arrays and linear processing," *IEEE J. Sel. Areas Commun.,* vol.32, no.9, pp.1721–1737, June 2014.

[26] D. Bethanabhotla, O.Y. Bursalioglu, H.C. Papadopoulos, and G. Caire, "Optimal user-cell association for massive MIMO wireless networks," *IEEE Trans. Wireless Commun.,* vol.15, no.3, pp.1835–1850, Mar. 2016.

[27] Q. Ye, O.Y. Bursalioglu, H.C. Papadopoulos, C. Caramanis, and J.G. Andrews, "User association and interference management in massive MIMO hetNets," *IEEE Trans. Wireless Commun.,* vol.64, no.5, pp.2049–2065, May 2016.

[28] F. Fernandes, A. Ashikhmin, and T.L. Marzetta, "Inter-cell interference in noncooperative TDD large scale antenna systems," *IEEE J. Select. Areas Commun.,* vol.31, no.2, pp.192–201, Feb. 2013.

[29] B. Akgun, M. Krunz, and O.O. Koyluoglu, "Pilot contamination attacks in massive MIMO systems," in *Proc. of the IEEE CNS'17,* Las Vegas, NV, Oct. 2017.

[30] N. Wang, E. Hossain, and V.K. Bhargava, "Joint downlink cell association and bandwidth allocation for wireless backhauling in two-tier hetNets with large-scale antenna arrays," *IEEE Trans. Wireless Commun.,* vol.15, no.5, pp.3251–3268, May 2016.

[31] Q. Ye, B. Rong, Y. Chen, M.A.-Shalash, C. Caramanis, and J.G. Andrews, "User association for load balancing in heterogeneous cellular networks," *IEEE Trans. Wireless Commun.,* vol.12, no.6, pp.2706–2716, June 2013.

[32] X. Li, T. Jiang, S. Cui, J. An, and Q. Zhang, "Cooperative communications based on rateless network coding in distributed MIMO systems [Coordinated and Distributed MIMO]," *IEEE Wireless Commun.,* vol.17, no.3, pp. 60–67, June 2010.

[33] O. Arnold, F. Richter, G. Fettweis, and O. Blume, "Power consumption modeling of different base station types in heterogeneous cellular networks," in *Proc. Future Netw. Mobile Summit,* Florence, Italy, June 2010, pp.1–8.

[34] Y. Huang, et al., "Energy-efficient design in heterogeneous cellular networks based on large-scale user behavior constraints," *IEEE Trans. Wireless Commun.,* vol.13, no.9, pp.4746–4757, Sept. 2014.

[35] L. Chen, et al., "Green full-duplex self-backhaul and energy harvesting small cell networks with massive MIMO," *IEEE J. Sel. Areas Commun.,* vol.34, no.12, pp.3709–3724, Dec. 2016.

[36] H. Rahbari, P. Siyari, M. Krunz, and J. Park, "Adaptive demodulation for wireless systems in the presence of frequency-offset estimation errors," in *Proc. IEEE INFOCOM'18,* Honolulu, HI, Apr. 2018.

[37] E. Oh, K. Son, and B. Krishnamachari, "Dynamic base station switching-on/off strategies for green cellular networks," *IEEE Trans. Wireless Commun.,* vol.12, no.5, pp.2126–2136, May 2013.

[38] M. Feng, S. Mao, and T. Jiang, "Dynamic base station sleep control and RF chain activation for energy efficient millimeter wave cellular systems," *IEEE Transactions on Veh. Tech.,* Under review.

[39] X. Guo, Z. Niu, S. Zhou, and P.R. Kumar, "Delay-constrained energy-optimal base station sleeping control," *IEEE J. Sel. Areas Commun.,* vol.34, no.5, pp.1073–1085, May 2016.

[40] 3GPP TS 36.420, "Evolved universal terrestrial radio access network (EUTRAN); X2 general aspects and principles," Dec. 2008.

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

[42] K. Xiao, S. Mao, and J.K. Tugnait, "Congestion control for infrastructure-based CRNs: A multiple model predictive control approach," in *Proc. IEEE GLOBECOM'2016,* Washington DC, Dec. 2016, pp. 1–6.

[43] W. Yu, "Multiuser water-flling in the presence of crosstalk," in *Proc. IEEE ITA Workshop*, San Diego, CA, Jan. 2007, pp.414–420.

[44] A. Schrijver, *Theory of Linear and Integer Programming,* John Wiley & Sons, June 1998.

[45] C. Berenstein and R. Gay, *Complex Variables: An Introduction,* Springer, 1997.

[46] K. Xiao, S. Mao, and J.K. Tugnait, "QoE-driven resource allocation for DASH over OFDMA networks," in *Proc. IEEE GLOBECOM'2016,* Washington DC, Dec. 2016, pp. 1–6.

[47] Y. Xiao, M. Hirzallah, and M. Krunz, "Optimizing inter-operator network slicing over licensed and unlicensed bands," in *Proc. IEEE SECON'18, Hong Kong,* June 2018.

[48] J. Xu, et al., "Cooperative distributed optimization for the hyper-dense small cell deployment," *IEEE Commun. Mag.,* vol. 52, no. 5, pp. 61–67, May 2014.

[49] Y.S. Soh, T.Q.S. Quek, M. Kountouris, and H. Shin, "Energy efficient heterogeneous cellular networks," *IEEE J. Select. Areas Commun.,* vol.31, no.5, pp.840–850, May 2013.

[50] I. Ashraf, L.T.W. Ho, and H. Claussen, "Improving energy efficiency of femtocell base stations via user activity detection," in *Proc. WCNC'10,* Sydney, Austrilia, Apr. 2010, pp.1–5.

[51] M. Feng and S. Mao, "Interference management in massive MIMO HetNets: A nested array approach," in *Proc. IEEE GLOBECOM'2016,* Washington DC, Dec. 2016, pp. 1–6.

[52] K. Hosseini, J. Hoydis, S. ten Brink, and M. Debbah, "Massive MIMO and small cells: How to densify heterogeneous networks," in *Proc. IEEE ICC'13,* Budapest, Hungary, June 2013, pp. 5442–5447.

[53] N. Wang, E. Hossain, and V. K. Bhargava, "Joint downlink cell association and bandwidth allocation for wireless backhauling in two-tier hetNets with large-scale antenna arrays," *IEEE Trans. Wireless Commun.,* vol. 15, no. 5, pp. 3251–3268, May 2016.

[54] M. Feng, T. Jiang, D. Chen, and S. Mao, "Cooperative small cell networks: High capacity for hotspots with interference mitigation," *IEEE Wireless Commun.,* vol. 21, no. 6, pp. 108–116, Dec. 2014.

[55] P. Pal and P. P. Vaidyanathan, "Nested arrays: A novel approach to array processing with enhanced degrees of freedom," *IEEE Trans. Sig. Proc.,* vol. 58, no. 8, pp. 4167–4180, Aug. 2010.

[56] R. T. Hoctor and S. A. Kassam, "The unifying role of the coarray in aperture synthesis for coherent and incoherent imaging," *Proc. IEEE,* vol. 78, no. 4, pp. 735–752, Apr. 1990.

[57] R. Gomory, "Outline of an algorithm for integer solutions to linear programs," *Bull. Amer. Math. Soc.,* vol. 64, no. 5, pp. 275–278, Sept. 1958.

[58] R. W. Irving, "An efficient algorithm for the 'Stable Roommates' problem," *Journal of Algorithms,* vol. 6, no. 6, pp.5 77–595, Dec. 1985.

[59] International Telecommunication Union, *Guidelines for Evaluation of Radio Transmission Technologies for IMT-2000*, Recommendation ITU-R M.1225, 1997.

[60] F. Boccardi, R. W. Heath, A. Lozano, T. L. Marzetta, and P. Popovski, "Five disruptive technology directions for 5G," *IEEE Commun. Magazine*, vol. 52, no. 2, pp. 74–80, February 2014.

[61] F. Rusek, D. Persson, B. Lau, E. Larsson, T. Marzetta, O. Edfors, and F. Tufvesson, "Scaling up MIMO: Opportunities and challenges with very large arrays," *IEEE Sig. Proc. Mag.,* vol. 30, Jan. 2013, pp. 40–60.

[62] Y. Huang, C. W. Tan, and B. D. Rao, "Joint beamforming and power control in coordinated multicell: Max-min duality, effective network and large system transition," *IEEE Trans. Wireless Commun.,* vol. 12, no. 6, pp. 2730–2742, June 2013.

[63] R. Zakhour and S. V. Hanly, "Base station cooperation on the downlink: Large system analysis," *IEEE Trans. Inf. Theory,* vol. 58, no. 4, pp. 2079–2106, Apr. 2012.

[64] J. Zuo, J. Zhang, C. Yuen, W. Jiang, and W. Luo, "Energy-efficient downlink transmission for multicell massive DAS with pilot contamination," *IEEE Trans. Veh. Technol.,* vol. 66, no. 2, pp. 1209–1221, Feb. 2017.

[65] X. Guo, S. Chen, J. Zhang, X. Mu, and L. Hanzo, "Optimal pilot design for pilot contamination elimination/reduction in large-scale multiple-antenna aided OFDM systems," *IEEE Trans. Wireless Commun.,* vol. 15, no. 11, pp. 7229–7243, Nov. 2016.

[66] X. Rao and V. K. N. Lau, "Compressive sensing with prior support quality information and application to massive MIMO channel estimation with temporal correlation," *IEEE Trans. Sig. Proc.,* vol. 63, no. 18, pp. 4914–4924, Sept. 2015.

[67] H. Xie, F. Gao, and S. Jin,"An overview of low-rank channel estimation for massive MIMO systems," *IEEE Access,* vol. 4, pp. 7313–7321, Nov. 2016.

[68] D. Kong, D. Qu, K. Luo, and T. Jiang, "Channel estimation under staggered frame structure for massive MIMO system," *IEEE Trans. Wireless Commun.,* vol. 15, no. 2, pp. 1469–479, Feb. 2016.

[69] Y. Zhu, L. Liu, A. Wang, K. Sayana, and J. Zhang, "DoA estimation and capacity analysis for 2D active massive MIMO systems," in *Proc. IEEE ICC'13,* Budapest, Hungary, June 2013, pp. 4630–4634.

[70] H. Yin, D. Gesbert, M. Filippou, and Y. Liu, "A coordinated approach to channel estimation in large-scale multiple-antenna systems," *IEEE J. Select. Areas Commun.,* vol.31, no.2, pp. 264–273, Feb. 2013.

[71] X. Wang, L. Gao, S. Mao, and S. Pandey, "DeepFi: Deep learning for indoor fingerprinting using channel state information," in *Proc. IEEE WCNC'15,* New Orleans, LA, Mar. 2015, pp. 1666–1671.

[72] X. Wang, L. Gao, S. Mao, and S. Pandey, "CSI-based fingerprinting for indoor localization: A deep learning approach," *IEEE Trans. Veh. Technol.,* vol. 66, no. 1, pp. 763–776, Jan. 2017.

[73] H. Xie, F. Gao, S. Zhang, and S. Jin, "A unified transmission strategy for TDD/FDD massive MIMO systems with spatial basis expansion model," *IEEE Trans. Veh. Technol.,* vol.66, no.4, pp. 3170–3184, Apr. 2017.

[74] J. Ma, S. Zhang, H. Li, N. Zhao, and A. Nallanathan, "Pattern division for massive MIMO networks with two-stage precoding," *IEEE Commun. Letters,* DOI: 10.1109/LCOMM.2017.2687868.

[75] K. Xiao, S. Mao, J.K. Tugnait, "Congestion control for infrastructure-based CRNs: A multiple model predictive control approach," *IEEE Trans. Wireless Commun.,* vol.16, no.4, pp.2614–2626, Apr. 2017.

[76] V. Chandrasekhar, J. G. Andrews, T. Muharemovic, Z. Shen, and A. Gatherer, "Power control in two-tier femtocell networks," *IEEE Trans. Wireless Commun.,* vol. 8, no. 8, pp. 4316–4328, Aug. 2009.

[77] M. Feng, D. Chen, Z. Wang and T. Jiang, "Throughput improvement for OFDMA femtocell networks through spectrum allocation and access control strategy," in *Proc. IEEE ComComAP'12,* Hong Kong, China, Jan. 2012, pp.387–391.

[78] W. C. Cheung, T. Q. S. Quek, and M. Kountouris, "Throughput optimization, spectrum allocation, and access control in two-tier femtocell networks," *IEEE J. Select. Areas Commun.,* vol. 33, no. 6, pp. 1171–1186, June 2015.

[79] D.-C. Oh, H.-C. Lee, and Y.-H. Lee, "Power control and beamforming for femtocells in the presence of channel uncertainty," *IEEE Trans. Veh. Technol.,* vol. 60, no. 6, pp. 2545–2554, July 2011.

[80] J. Xiang, Y. Zhang, T. Skeie, and L. Xie, "Downlink spectrum sharing for cognitive radio femtocell networks," *IEEE Systems J.*, vol. 4, no. 4, pp. 524–534, Dec. 2010.

[81] H. Zhou, S. Mao, and P. Agrawal, "Approximation algorithms for cell association and scheduling in femtocell networks," *IEEE Transactions on Emerging Topics in Computing,* vol. 3, no. 3, pp. 432-443, Sept. 2015.

[82] M. Feng, Z. He, and S. Mao, "QoE driven video streaming over cognitive radio networks for multi-user with single channel access," *IEEE ComSoc MMTC Communications-Frontiers,* vol.12, no.2, pp.7–11, Mar. 2017.

[83] M. Feng, S. Mao, and T. Jiang, "Joint frame design, resource allocation and user association for massive MIMO heterogeneous networks with wireless backhaul," *IEEE Trans. Wireless Commun.,* vol.17, no.3, pp.1937–1950, Mar. 2018.

[84] X. Ge, H. Cheng, M. Guizani, and T. Han, "5G wireless backhaul networks: Challenges and research advances," *IEEE Network,* vol.28, no.6, pp.6–11, Nov. 2014.

[85] U. Siddique, H. Tabassum, E. Hossain, and D. I. Kim, "Wireless backhauling of 5G small cells: Challenges and solution approaches," *IEEE Wireless Commun.,* vol.22, no.5, pp.22–31, Oct. 2015.

[86] N. Wang, E. Hossain, and V. K. Bhargava, "Backhauling 5G small cells: A radio resource management perspective," *IEEE Wireless Commun.,* vol.22, no.5, pp.41–49, Oct. 2015.

[87] N. Wang, E. Hossain, and V. K. Bhargava, "Joint downlink cell association and bandwidth allocation for wireless backhauling in two-tier HetNets with large-scale antenna arrays," *IEEE Trans. Wireless Commun.,* vol.15, no.5, pp.3251–3268, May 2016.

[88] B. Li, D. Zhu, and P. Liang, "Small cell in-band wireless backhaul in massive MIMO systems: A cooperation of next-generation techniques," *IEEE Trans. Wireless Commun.,* vol.14, no.12, pp.7057–7069, Dec. 2015.

[89] H. Tabassum, A. H. Sakr, E. Hossain, "Analysis of massive MIMO-enabled downlink wireless backhauling for full-duplex small cells," *IEEE Trans. Commun.,* vol.64, no.6, pp.2354–2369, June 2016.

[90] Z. Gao, L. Dai, D. Mi, Z. Wang, M. A. Imran, and M. Z. Shakir, "MmWave massive-MIMO-based wireless backhaul for the 5G ultra-dense network," *IEEE Wireless Commun.,* vol.22, no.5, pp.13–21, Oct. 2015.

[91] H. Kim and Y. Han, "A proportional fair scheduling for multicarrier transmission systems," *IEEE Commun. Lett.,* vol.9, no.3, pp.210–212, Mar. 2005.

[92] X. Wang, L. Gao, and S. Mao, "CSI phase fingerprinting for indoor localization with a deep learning approach," *IEEE Internet of Things,* vol.3, no.6, pp.1113–1123, Dec. 2016.

[93] W. Wang, Y. Chen, Q. Zhang, T. Jiang, "A software-defined wireless networking enabled spectrum management architecture," *IEEE Commun. Mag.,* vol.54, no.1, pp. 33–39, Jan. 2016.

[94] D. P. Palomar and M. Chiang, "A tutorial on decomposition methods for network utility maximization," *IEEE J. Sel. Areas Commun.,* vol.24, no.8, pp.1439–1451, Aug. 2006.

[95] E. G. Larsson, F. Tufvesson, O. Edfors, and T. L. Marzetta, "Massive MIMO for next generation wireless systems," *IEEE Commun. Mag.,* vol.52, no.2, pp.186–195, Feb. 2014.

[96] L. Lu, G. Y. Li, A. L. Swindlehurst, A. Ashikhmin, and R. Zhang, "An overview of massive MIMO: benefits and challenges," *IEEE J. Sel. Topics Signal Processing,* vol.8, no.5, pp. 742–758, Oct. 2014.

[97] S. Samarakoon, M. Bennis, W. Saad, M. Ltva-aho, "Backhaul-aware interference management in the uplink of wireless small cell networks," *IEEE Trans. Wireless Commun.,* vol.12, no.11, pp. 5813–5825, Nov. 2013.

[98] L. Sanguinetti, A. L. Moustakas, and M. Debbah, "Interference management in 5G reverse TDD hetNets with wireless backhaul: A large system analysis," *IEEE J. Select. Areas Commun.,* vol.33, no.6, pp. 1187–1200, June 2015.

[99] H. H. Yang, G. Geraci, T. Q. S. Quek, "Energy-efficient design of MIMO heterogeneous networks with wireless backhaul," *IEEE Trans. Wireless Commun.,* vol. 15, no. 7, pp. 4914–4927, July 2016.

[100] D. Liu, L. Wang, Y. Chen, M. Elkashlan, K.-K. Wong, R. Schober and L. Hanzo, "User association in 5G networks: A survey and an outlook," *IEEE Commun. Surveys Tuts.,* vol.18, no.2, pp. 1018–1044, second quarter 2016.

[101] Y. Lin, W. Bao, W, Yu, B. Liang, "Optimizing user association and spectrum allocation in HetNets: A utility perspective," *IEEE J. Select. Areas Commun.,* vol.33, no.6, pp. 1025–1039, June 2015.

[102] S. Mao and S.S. Panwar, "A survey on envelope processes and their applications in quality of service provisioning," *IEEE Communications Surveys and Tutorials,* vol.8, no.3, pp.2–19, Third Quarter, 2006.

[103] M. Feng, S. Mao, and T. Jiang, "Enhancing the performance of future wireless networks with Software Defined Networking," *Springer Frontiers of Information Technology and Electronic Engineering Journal,* vol.17, no.7, pp.606–619, July 2016.

[104] M. Feng, S. Mao, and T. Jiang, "Duplex mode selection and channel allocation for full-duplex cognitive femtocell networks," in *Proc. IEEE WCNC'15,* New Orleans, LA, Mar. 2015, pp.1900–1905.

[105] M. Feng, S. Mao, and T. Jiang, "Joint duplex mode selection, channel allocation, and power control for full-duplex cognitive femtocell networks," *Elsevier Digital Communications and Networks,* vol.1, no.1, pp.30–44, Feb. 2015.

[106] L. Li, J.P. Seymour, L.J. Cimini, and C.C. Shen, "Coexistence of Wi-Fi and LAA networks with adaptive energy detection," *IEEE Trans. Veh. Technol.,* vol.66, no.11, pp. 10384-10393, Nov. 2017.

[107] S.-M. Cheng, S.-Y. Lien, F.-S. Chu, and K.-C. Chen, "On exploiting cognitive radio to mitigate interference in macro/femto heterogeneous networks," *IEEE Wireless Commun.,* vol.18, no.3, pp.40–47, June 2011.

[108] L. Huang, G. Zhu, and X. Du, "Cognitive femtocell networks: An opportunistic spectrum access for future indoor wireless coverage," *IEEE Wireless Commun.,* vol.20, no.2, pp.41–51, Apr. 2013.

[109] R. Xie, F. R. Yu, H. Ji, and Y. Li, "Energy-efficient resource allocation for heterogeneous cognitive radio networks with femtocells," *IEEE Trans. Wireless Commun.,* vol.11, no.11, pp.3910–3920, Nov. 2012.

[110] S.-Y. Lien, Y.-Y. Lin, and K.-C. Chen, "Cognitive and game-theoretical radio resource management for autonomous femtocells with QoS guarantees," *IEEE Trans. Wireless Commun.,* vol.10, no.7, pp.2196–2206, June 2011.

[111] J. Huang and V. Krishnamurthy, "Cognitive base stations in LTE/3GPP femtocells: A correlated equilibrium game-theoretic approach," *IEEE Trans. Commun.,* vol.59, no.12, pp.3485–3493, Oct. 2011.

[112] C. Long, Y. Cao, T. Jiang, and Q. Zhang, "Edge computing framework for cooperative video processing in multimedia IoT systems," *IEEE Trans. Multimedia,* vol.20, no.5, pp.1126–1139, May 2018.

[113] W. Afifi and M. Krunz, "TSRA: An adaptive mechanism for switching between communication modes in full-duplex opportunistic spectrum access systems," *IEEE Transactions on Mobile Computing,* vol.16, no.6, pp. 1758–1772, June 2017.

[114] J. Choi, M. Jain, K. Srinivasan, P. Levis, and S. Katti, "Achieving single channel, full duplex wireless communication," in *Proc. ACM MOBICOM'10,* Chicago, IL, 2010, pp. 1–12.

[115] S. Goyal, P. Liu, S. Hua, and S. Panwar, "Analyzing a full-duplex cellular system," in *Proc. IEEE CISS'13,* Baltimore, MD, 2013, pp. 1–6.

[116] L. Li, L.J. Cimini and Y. Xiao, "Spectral efficiency of cooperative full-duplex relaying with imperfect channel estimation," in *Proc. IEEE GLOBECOM'14,* Austin, TX, Dec. 2014, pp.4203–4208.

[117] W. Wang and Q. Zhang, "Local cooperation architecture for self-healing femtocell networks," *IEEE Wireless Commun. Mag.,* vol.21, no.2, pp. 42–49, Apr. 2014.

[118] S. Saadat, D. Chen, K. Luo, M. Feng, and T. Jiang, "License assisted access-WiFi coexistence with TXOP backoff for LTE in unlicensed band," *China Communications,* vol.14, no.3 pp.1–14, Apr. 2017.

[119] P. Siyari, M. Krunz, and D. Nguyen, "Power games for secure communications in single-stream MIMO interference networks," *IEEE Trans. Wireless Commun.,* to appear.

[120] J. Papandriopoulos and J. Evans, "Low-complexity distributed algorithms for spectrum balancing in multi-user DSL networks," in *Proc. IEEE ICC'06,* Istanbul, Turkey, 2006, pp. 3270–3275.

[121] L. Venturino, N. Prasad, and X. Wang, "Coordinated scheduling and power allocation in downlink multicell OFDMA networks," *IEEE Trans. Veh. Technol.,* vol.58, no.6, pp.2835–2848, Jan. 2009.

[122] W. Cheng, X. Zhang, and H. Zhang, "Optimal dynamic power control for full-duplex bidirectional-channel based wireless networks," in *Proc. IEEE INFOCOM'13,* Turin, Italy, 2013, pp. 3120–3128.

[123] L. Li, L.J. Cimini, and X.G. Xia, "Impact of direct link on outage of cooperative full-duplex relaying," in *Proc. IEEE CISS'15,* Baltimore, MD, Mar. 2015, pp.1–6.

[124] C. Ni, M. Feng, K. Luo, T. Jiang, and S. Mao, "Additive cancellation signal method for sidelobe suppression in NC-OFDM based cognitive radio systems," in *Proc. IEEE GLOBECOM 2015,* San Diego, CA, Dec. 2015, pp.1–5.

[125] X. Wang, L. Gao, and S. Mao, "PhaseFi: Phase fingerprinting for indoor localization with a deep learning approach," in *Proc. IEEE GLOBECOM'15*, San Diego, CA, Dec. 2015, pp.1–6.

[126] R. Irving, "An efficient algorithm for the 'Stable Roommates' problem," *Journal of Algorithms,* vol.6, no.6, pp.577–595, 1985.

[127] L. Li, A. Song, L.J. Cimini, X.-G. Xia, and C.-C. Shen, "Interference cancellation in in-band full-duplex underwater acoustic systems," in *Proc. IEEE OCEANS'15,* Washington DC, Oct. 2015, pp. 1–6.

[128] A. T. Hoang and Y.-C. Liang, "Downlink channel assignment and power control for cognitive radio networks," *IEEE Trans. Wireless Commun.,* vol.7, no.8, pp.3106–3117, Aug. 2008.

[129] P. Gao, D. Chen, M. Feng, D. Qu, and T. Jiang, "On the interference avoidance method in two-tier LTE networks with femtocells," in *Proc. IEEE WCNC'13,* Shanghai, China, Apr. 2013, pp.3585–3590.

[130] M. Feng, D. Chen, Z. Wang, T. Jiang, and D. Qu, "An improved spectrum management scheme for OFDMA femtocell networks," in *Proc. IEEE ICCC 2012,* Beijing, China, Aug. 2012, pp.132–136.

[131] Y. Xing, C. N. Mathur, M. Haleem, R. Chandramouli, and K. Subbalakshmi, "Dynamic spectrum access with QoS and interference temperature constraints," *IEEE Trans. Mobile Comput.,* vol.6, no.4, pp.423–433, Feb. 2007.

[132] A. T. Hoang, Y.-C. Liang, and M. H. Islam, "Power control and channel allocation in cognitive radio networks with primary users' cooperation," *IEEE Trans. Mobile Comput.,* vol.9, no.3, pp.348–360, Mar. 2010.

[133] G. Zheng, I. Krikidis, and B. Ottersten, "Full-duplex cooperative cognitive radio with transmit imperfections," *IEEE Trans. Wireless Commun.,* vol.12, no.5, pp.2498–2511, Apr. 2013.

[134] H. Kim, S. Lim, H. Wang, and D. Hong, "Optimal power allocation and outage analysis for cognitive full duplex relay systems," *IEEE Trans. Wireless Commun.,* vol.11, no.10, pp.3754–3765, Sept. 2012.

[135] I. Krikidis, H. A. Suraweera, P. J. Smith, and C. Yuen, "Full-duplex relay selection for amplify-and-forward cooperative networks," *IEEE Trans. Wireless Commun.,* vol.11, no.12, pp.4381–4393, Oct. 2012.

[136] W. Cheng, X. Zhang, and H. Zhang, "Full/half duplex based resource allocations for statistical quality of service provisioning in wireless relay networks," in *Proc. IEEE INFOCOM'12,* Orlando, FL, 2012, pp. 864–872.

[137] T. Riihonen, S. Werner, and R. Wichman, "Hybrid full-duplex/half-duplex relaying with transmit power adaptation," *IEEE Trans. Wireless Commun.,* vol.10, no.9, pp.3074–3085, July 2011.

[138] I.C.-Lin, L. J. Greenstein, and R. D. Gitlin, "A macrocell/microcell cellular architecture for low- and high-mobility wireless users," *IEEE J. Sel. Areas Commun.,* vol. 11, no. 6, pp. 885–891, Aug. 1993.

[139] J. Hoydis, M. Kobayashi, and M. Debbah, "Green small-cell networks," *IEEE Veh. Technol. Mag.,* vol.6, no.1, pp.37–43, Mar. 2011.

[140] T.S. Rappaport, *Wireless Communications: Principles & Practice,* Upper Saddle River, NJ, Prentice Hall PTR, 1999.

[141] T.-S. P. Yum and W.-S. Wong, "Hot-spot traffic relief in cellular systems," *IEEE J. Sel. Areas Commun.,* vol.11, no.6, pp.934–940, Aug. 1993.

[142] W. Song and W. Zhuang and Y. Cheng, "Load balancing for cellular/WLAN integrated networks," *IEEE Netw.,* vol. 21, no. 1, pp.6–12, Jan./Feb. 2007.

[143] V. Chandrasekhar, J.G. Andrews, and A. Gatherer, "Femtocell networks: a survey," *IEEE Commun. Mag.,* vol. 46, no. 9, pp.59–67, Sep. 2008.

[144] A. Saleh, A. Rustako, and R. Roman, "Distributed antennas for indoor radio communications," *IEEE Trans. Commun.,* vol.35, no.12, pp. 1245–1251, Dec. 1987.

[145] P. Ralf, *et al.*, "Relay-based deployment concepts for wireless and mobile broadband radio," *IEEE Commun. Mag.,* vol.42, no.9, pp. 80–89, Sept. 2004.

[146] H.-C. Lu, and W. Liao, "Cooperative strategies in wireless relay networks," *IEEE J. Sel. Areas Commun.,* vol.30, no.2, pp. 323–330, Feb. 2012.

[147] R.R. Choudury, X. Yang, R. Ramanathan and N. Vaidya, "On designing MAC protocols for wireless networks with directional antennas," *IEEE Trans. Mobile Comput.,* vol.5, no.5, pp.477–491, May 2006.

[148] J. Mitola, and G.Q. Maguire, "Cognitive radio: making software radios more personal," *IEEE Personal Commun. Mag.,* vol.6, no.4, pp.13–18, Aug. 1999.

[149] V.K. Garg, and J.E. Wilkes, *Wireless and Personal Communications Systems,* Upper Saddle River, NJ, Prentice Hall PTR, 1996.

[150] Y.I. Choni and A. Hassan, "Optimal adaptive antenna arrays for asynchronous communication systems," *IEEE Trans. Antennas and Propagation,* vol.60, no.6, pp.3071–3076, June 2012.

[151] J.-S. Wu, J.-K. Chung, and C.-C. Wen, "Hot-spot traffic relief with a tilted antenna in CDMA cellular networks," *IEEE Trans. Veh. Technol.,* vol.47, no.1, pp.1–9, Feb. 1998.

[152] M. Feng, S. Mao, and T. Jiang, "Dealing with link blockage in mmWave networks: D2D relaying or multi-beam reflection?," in *Proc. IEEE PIMRC'17*, Montreal, Canada, Oct. 2017.

[153] T. S. Rappaport, et al., "Millimeter wave mobile communications for 5G cellular: It will work!" *IEEE Access J.*, vol. 1, May 2013, pp. 335–449.

[154] A. Alkhateeb, G. Leus, and R.W. Heath, "Limited feedback hybrid precoding for multi-user millimeter wave systems," *IEEE Trans. Wireless Commun.,* vol.14, no.11, pp.6481–6494, Nov. 2015.

[155] Z. He and S. Mao, "A decomposition principle for link and relay selection in dual-hop 60 GHz networks," in *Proc. IEEE INFOCOM'16*, San Francisco, CA, Apr. 2016, pp.1683–1691.

[156] Z. He, S. Mao, S. Kompella, and A. Swami, "On link scheduling in dual-hop 60 GHz mmWave networks," *IEEE Trans. Vehicular Technology*, to appear. DOI: 10.1109/TVT.2017.2717840.

[157] M. Feng and S. Mao, "Dealing with limited backhaul in millimeter wave systems: A deep reinforcement learning approach," *IEEE Commun. Mag.,* under review.

[158] J. Qiao, X. Shen, J.W. Mark, Q. Shen, Y. He, and L. Lei, "Enabling device-to-device communications in millimeter-wave 5G cellular networks," *IEEE Commun. Mag.,* vol.53, no.1, pp.209–215, Jan. 2015.

[159] Y. Cao, C. Long, T. Jiang, and S. Mao, "Share communication and computation resources on mobile devices: A social-awareness perspective," *IEEE Wireless Commun.,* vol.23, no.4, pp.52–59, Aug. 2016.

[160] Y. Niu, C. Gao, Y. Li, L. Su, D. Jin, and A.V. Vasilakos, "Exploiting device-to-device communications in joint scheduling of access and backhaul for mmWave small cells," *IEEE J. Sel. Areas Commun.,* vol.33, no.10, pp.2052–2069, Oct. 2015.

[161] S. Mao, S. Lin, S.S. Panwar, Y. Wang, and E. Celebi, "Video transport over ad hoc networks: multistream coding with multipath transport," *IEEE J. Select. Areas Commun.,* vol.21, no.10, pp. 1721–1731, Dec. 2003.

[162] Q. Xue, X. Fang, and C.-X. Wang, "Beamspace SU-MIMO for future millimeter wave wireless communications," *IEEE J. Sel. Areas Commun.,* vo.35, no.7, pp.1564–1575, July 2017.

[163] S.-C. Lin and I.F. Akyildiz, "Dynamic base station formation for solving NLOS problem in 5G millimeter-wave communication," in *Proc. IEEE INFOCOM'17,* Atlanta, GA, May 2017.

[164] I.K. Son, S. Mao, M.X. Gong, and Y. Li, "On frame-based scheduling for directional mmWave WPANs," in *Proc. IEEE INFOCOM'12,* pp.2149–2157, Orlando, FL, March 2012.

[165] J.G. Andrews, et al., "Modeling and analyzing millimeter wave cellular systems," *IEEE Trans. Commun.,* vol.65, no.1, pp.6481–6494, Jan. 2017.

[166] Y. Wang, S. Mao, and T.S. Rappaport, "On directional neighbor discovery in mmWave networks," in *Proc. IEEE ICDCS'17,* Atlanta, GA, June 2017, pp.1704–1713.

[167] J. Qiao, et al., "Enabling multi-hop concurrent transmissions in 60 GHz wireless personal area networks," *IEEE Trans. Wireless Commun.,* vol.10, no.11, pp.3824–3833, Nov. 2011.

[168] Z. He, S. Mao, and T.S. Rappaport, "On link scheduling under blockage and interference in 60 GHz ad hoc networks," *IEEE Access J.*, vol.3, pp.1437–1449, Sept. 2015.

Appendices

Appendix A

Publications

A.1 Conference Publications

1. **Mingjie Feng**, Shiwen Mao, and Tao Jiang, Duplex mode selection and channel allocation for full-duplex cognitive femtocell networks, in *Proc. IEEE WCNC 2015*, New Orleans, LA, Mar. 2015, pp.1900-1905.

2. Chunxing Ni, **Mingjie Feng**, Kai Luo, Tao Jiang, and Shiwen Mao, Additive cancellation signal method for sidelobe suppression in NC-OFDM based cognitive radio systems, in *Proc. IEEE GLOBECOM 2015*, San Diego, CA, Dec. 2015, pp.1-5.

3. **Mingjie Feng**, Shiwen Mao, and Tao Jiang, BOOST: Base station on-off switching strategy for energy efficient massive MIMO HetNets, in *Proc. IEEE INFOCOM 2016*, San Francisco, CA, Apr. 2016, pp.1395-1403.

4. **Mingjie Feng** and Shiwen Mao, Interference management in massive MIMO HetNets: A nested array approach, in *Proc. IEEE GLOBECOM 2016*, Washington DC, Dec. 2016, pp.1-6.

5. **Mingjie Feng** and Shiwen Mao, Adaptive pilot design for massive MIMO hetNets with wireless backhaul, in *Proc. IEEE SECON 2017*, San Diego, CA, June 2017, pp.1-9.

6. **Mingjie Feng**, Shiwen Mao, and Shiwen Mao, Dealing with link blockage in mmWave networks: D2D relaying or multi-beam reflection?, in *Proc. IEEE PIMRC 2017*, Montreal, Canada, Oct. 2017, pp.1-5.

A.2 Journal Publications

1. **Mingjie Feng**, Tao Jiang, Da Chen, and Shiwen Mao, Cooperative small cell networks: High capacity for hotspots with interference mitigation, *IEEE Wireless Communications*, vol.21, no.6, pp.108-116, Dec. 2014.

2. **Mingjie Feng**, Shiwen Mao, and Tao Jiang, Joint duplex mode selection, channel allocation, and power control for full-duplex cognitive femtocell networks, *Elsevier Digital Communications and Networks Journal*, vol.1, no.1, pp.30-44, Feb. 2015.

3. **Mingjie Feng** and Shiwen Mao, Harvest the potential of massive MIMO with multi-layer techniques, *IEEE Network*, vol.30, no.5, pp.40-45, Sept./Oct. 2016.

4. **Mingjie Feng**, Shiwen Mao, and Tao Jiang, Enhancing the performance of future wireless networks with Software Defined Networking, *Springer Frontiers of Information Technology and Electronic Engineering Journal*, vol.17, no.7, pp.606-619, July 2016.

5. **Mingjie Feng**, Zhifeng He, and Shiwen Mao, QoE driven video streaming over cognitive radio networks for multi-user with single channel access, *IEEE ComSoc MMTC Communications-Frontiers*, vol.12, no.2, pp.7-11, Mar. 2017.

6. Salman Saadat, Da Chen, Kai Luo, **Mingjie Feng**, and Tao Jiang, License assisted access-WiFi coexistence with TXOP backoff for LTE in unlicensed band, *China Communications*, vol.14, no.3, pp.1-14, Apr. 2017.

7. **Mingjie Feng**, Shiwen Mao, and Tao Jiang, Base station ON-OFF switching in 5G wireless networks: Approaches and challenges, *IEEE Wireless Communications*, vol.24, no.4, pp.46-54, Aug. 2017.

8. **Mingjie Feng**, Shiwen Mao, and Tao Jiang, BOOST: Base station on-off switching strategy for green massive MIMO HetNets, *IEEE Transactions on Wireless Communications*, vol.16, no.11, pp.7319-7332, Nov. 2017.

9. **Mingjie Feng** and Shiwen Mao, Interference management and user association for nested array-based massive MIMO HetNets, *IEEE Transactions on Vehicular Technology*, vol.67, no.1, pp.454-466, Jan. 2018.

10. **Mingjie Feng**, Shiwen Mao, and Tao Jiang, Joint frame design, resource allocation, and user association for massive MIMO heterogeneous networks with wireless backhaul, *IEEE Transactions on Wireless Communications*, vol.17, no.3, pp.1937-1950, Mar. 2018.