QUEUE-AWARE PERFORMANCE OPTIMIZATION OF HETEROGENEOUS NETWORKS

by

Fancheng Kong

Dissertation submitted in fulfilment of the requirements

for the degree of

DOCTOR OF PHILOSOPHY

School of Computing & Communications
Faculty of Science
University of Technology Sydney
Sydney, Australia

August 24, 2018

ABSTRACT

To meet surging traffic demands, heterogeneous networks (HetNets) enable a more flexible, targeted and economical deployment of new infrastructure versus tower-mounted macro-only systems, which are very expensive to deploy and maintain. For a high network spectrum efficiency, load balancing across different tiers can be achieved by optimizing the association between users and base stations (BSs). To achieve a high energy efficiency, proper controls of BSs' activation (on/off status) and deployment density can significantly avoid unnecessary BS power consumption. However, some practical conditions are not considered in existing studies.

(i) Previous studies usually assumed that BSs were always busy transmitting packets to their associated users, which characterized a worst case of the performance metrics. In practice, one BS can either be busy or idle, depending on its queuing condition, in which case the performance metrics such as the packet delay should be further studied with queuing taken into account.

(ii) With the assumption of continuous BS transmission in previous literatures, the network power consumption linearly increases with the number of BSs only. Practically, the power consumption of a BS in the idle state is much lower than that in the busy state, the tuning of the network design parameters, for example, the bandwidth allocation and the BS deployment density, thus have a significant impact on the BS busy/idle status, which in turn affects the network energy efficiency.

(iii) Most of the previous studies focus on a uniform user distribution. In reality, users might not be evenly distributed and may form a cluster in certain

hot area. In such cases, the user association optimization in a per-tier fashion would result in a poor user experience in the overloaded areas, and a per-station association scheme is thus preferable.

To address the above considerations, the thesis focuses on the optimization of both the network spectrum efficiency and the network energy efficiency with practical assumptions of queuing and non-uniform user distribution, which is elaborated in the following.

1) Delay-optimal biased user association in HetNets. A thinned Poisson point process model to characterize the locations of BSs in the busy state, and an explicit expression of the average traffic intensity of each tier is obtained. On that basis, an optimization problem is formulated to minimize the lower bound of the network mean queuing delay by tuning the biasing factor of each tier, which is shown to be a convex problem. The simulation results demonstrate that the network queuing performance can be significantly improved by properly tuning the biasing factor. It is further shown that the network mean queuing delay might be improved at the cost of a deterioration of the network signal-to-interference ratio (SIR) coverage, which indicates a performance tradeoff between real-time and non-real-time traffic in HetNets.

2) Queue-aware optimal bandwidth allocation in HetNets. Based on the queuing analysis, a minimization problem of the network average power consumption and a maximization problem of the network SIR coverage are formulated, which are shown to be convex and concave with respect to the bandwidth allocation to each tier, respectively. By using an approximation of the average traffic intensity, closed-form solutions are obtained for both problems. Simulation results of a 2-tier HetNet demonstrate that the network average power consumption and the SIR coverage can be significantly improved by the optimal bandwidth allocation.

3) Queue-aware energy efficient base station density optimization in HetNets. By further using the approximation that BSs of a tier have the same SIR coverage, the cumulative distribution function (CDF) of the traffic intensity of each tier is obtained. On that basis, a minimization problem of the network average power consumption is studied by optimally tuning the activation ratio of micro BSs under the quality of serive (QoS) constraints of the network mean queuing delay and the network SIR coverage. Numerical results demonstrate that if the idle power coefficient is below a certain threshold, the optimal activation ratio should equal the one to minimize the network average power consumption per area. Otherwise, the optimal activation ratio should be obtained according to the QoS constraints. It is further shown that universal frequency reuse (UFR) outperforms spectrum partitioning (SP) in terms of both energy efficiency and SIR coverage in the considered scenario.

4) Optimal biased association scheme with non-uniform user distribution in HetNets. A practical scenario is studied where one cell is overloaded due to the cluster of users. By maximizing the mean user utility in the area of this overloaded cell and its neighboring cells, the optimal biasing factor can be obtained. It is found that in the scenario where the overloaded cell is fully surrounded by a macro cell, the optimal biasing factor logarithmically decreases with the user's intensity of the overloaded cell. Numerical results demonstrate that the mean user rate of the overloaded cell and the whole network can be significantly improved by properly tuning the biasing factor of the overloaded cell.

Key words: Heterogeneous network, Queuing, Non-uniform user distribution, Biasing factor, Bandwidth allocation, BS deployment density, Network mean queuing delay, Network average power consumption.

STATEMENT OF CANDIDATE

This thesis is the result of a research candidate conducted with another University as part of a collaborative Doctoral degree.

I certify that the work in this thesis has not previously been submitted for a degree nor has it been submitted as part of the requirements for a degree to any other university or institution other than University of Technology Sydney.

I also certify that the thesis is an original piece of research and it has been written by me.

In addition, I certify that all information sources and literature used are indicated in the thesis.

Production Note:
Signature removed prior to publication.
. . . . . .

Fancheng Kong

ACKNOWLEDGMENTS

I would like to express my profound gratitude to my principal supervisors, Professor Yingjie Jay Guo and Professor Hongbo Zhu, for their support, encouragement, supervision and guidance during this research work. Their vision, technical knowledge, moral support and continuous guidance helped me to complete my work successfully. They always stood with me whenever I encountered any difficulties during my postgraduate studies.

I also benefited from a number of people during my graduate studies. I am highly thankful to my co-supervisor, Associate Professor Xinghua Sun, for his research collaboration, guidance and valuable suggestions throughout this study. He has been and still is an inspiring personality for me. I should also like to thank Dr. Peiyuan Qin and Dr. Can Ding from Global Big Data Technology Centre for their guidance and support during this study. I owe my sincere thanks to Genie Tan for helping me with my enrollment to UTS and preparing everything upon my arrival.

I sincerely appreciate the financial support of the UTS President's scholarship and UTS top-up scholarship from Graduate Research School that made this study and my stay in Australia possible.

I express gratitude to my friends in Global Big Data Technology Centre. In particular, I would like to thank Dr. Shulin Chen, Dr. Haihan Sun and Dr. Yijiang Nan that helped make my stay enjoyable in Sydney

Finally, I would like to thank my mum Jie Wang and my dad Qinghai Kong who have always been proud of me and were always there for me. I could never ask for more. I should also like to thank my girlfriend Selena Liang, for always

supporting me when I was in tough times.

Contents

Chapter 1

Introduction

In this chapter, we discuss the research background of the thesis. The emergence of the heterogeneous networks (HetNets) is first introduced in Section 1.1. The key technologies in designing a HetNet in previous literatures are then reviewed in Section 1.2. The technical challenges and existing problems based on the literature review in HetNets are examined and the motivations of this thesis are elaborated in Section 1.3. At last, an overview of the research questions and the structure of the thesis are given in Section 1.4.

1.1 The Prevalence of HetNets

Future wireless networks are confronting tremendous demands by explosive number of subscribers and exponential growth in mobile data traffic [1]. According to Visual Network Index (VNI) report released from Cisco, mobile data traffic has grown 18-fold over the past 5 years [2] driven by smartphones, tablets, and video streaming. It is estimated that the wireless data explosion will continue to grow at a scale of 1000 times in 10 years [3]. Therefore, the data rate of the fifth generation (5G) should be enhanced by at least 1000 times to meet future communication demands. In addition, 5G is required to support massive accessed devices and diverse quality of service (QoS) as the number of devices

could reach the tens or even hundreds of billions by the time 5G comes to fruition [4–6].

As the long-term evolution (LTE) system embodying 4G has been widely deployed and is reaching maturity, under which case small amounts of new spectrum and limited performance improvements can be expected, a paradigm shift should be achieved in 5G via innovative new technologies [7]. To be more specific, the key technologies to get to 1000 times data rate in 5G can be categorized as:

1. Massive multiple-input multiple-output (MIMO) by implementing single BSs with hundreds of antennas to smooth out channel responses as all small-scale channel randomness abates as the number of channel observations grows [8–13];

2. Spectrum expansion by moving to millimeter wave (mmWave) spectrum to make a better use of WiFi's unlicensed spectrum [14–19];

3. Extreme densification by incorporating large number small-scale BSs such as femto-BS with traditional cellular BSs to improve area spectral efficiency [20–24].

Among them, the most straightforward but effective way to increase the network capacity is the extreme densification that makes the cells smaller [1]. It is predicted in [25] that in the not too distant future, say 10–15 years out, the number of BSs may actually exceed the number of cell phone subscribers, resulting in a cloud-like data shower where a mobile device may connect to multiple BSs. Through the massive deployment of various small-scale BSs, the network architecture is thus evolving to more dense and irregular heterogeneous networks (HetNets) [26–33].

HetNets enable a more flexible, targeted and economical deployment of new infrastructure versus tower-mounted macro-only systems, which are very expensive to deploy and maintain [34]. In a HetNet, various types of BSs are deployed to offload the macro cell users, forming a multi-tier network overlaid with many small cells. Fig. 1.1 illustrates

a 3-Tier HetNet which consists of macro BS, pico BS and femto BS, which differ primarily in terms of maximum transmit power, physical size, deployment density, and cost. For instance, the macro BSs are sparsely deployed and offer basic long-range coverage, while the widely deployed femto BSs can only provide short-range communication links to nearby users.

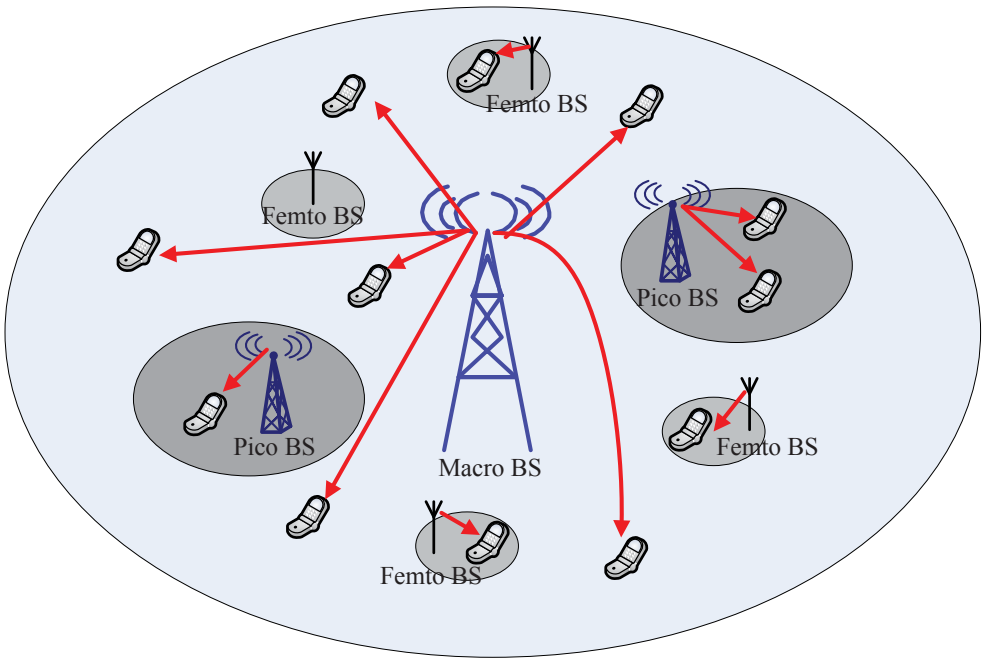


Figure 1.1: Illustration of a 3-Tier heterogeneous cellular network. Only a single macrocell is shown for simplicity.

However, the prevalence of HetNets have raised some challenges. First of all, as the transmission power of these small-scale BSs is usually 10–20dB lower than that of macro BSs, most of the users will still tend to associate with the macro BSs with the strongest downlink signal, which leads to the load imbalance across different tiers. Furthermore, the proliferation of small-scale BSs, nevertheless, leads to a significant increment on power consumption, which greatly raises the operation expenditure for service providers. Therefore, it is essential and necessary to optimize the design of a HetNet to improve its capacity while reduce the energy consumption. In the following, we will review the key technologies

to deal with these challenges.

1.2 Key Technologies in Designing HetNets

A great deal of effort has been made to improve the network performance. A large fraction of the previous studies on HetNets strived to optimize the network spectrum efficiency in terms of the rate and signal-to-interference-plus-noise ratio (SINR). In addition, some literatures placed their emphasis on the users' quality of service (QoS) provisioning by the performance metrics of the packet transmission delay. Recently, as the issue of global warming and heightened concern for the environment has raised a special focus on the energy efficiency in communication systems, more and more researchers have paid much attention to study power saving in HetNets. Therefore, Section 1.2.1 will review the literatures to optimize the spectrum efficiency, Section 1.2.2 will go through the studies on delay optimization, and Section 1.2.3 will summarize the related works on energy efficiency optimization in Hetnets.

1.2.1 Review of Spectrum Efficiency Optimization in HetNets

As mentioned before, in HetNets, the BS parameters such as transmission power and deployment density are distinct across tiers. Due to the disparate transmit powers and BS capabilities, mobile users would be much more likely to associate with a tower-mounted macro BS by the decision metrics such as SINR or received signal strength indicator (RSSI). Therefore, a conservative offloading approach may result in severe load imbalance [30], which would not only underutilize the benefit from the deployment of small-scale cells but would also deteriorate the multimedia performance of macro cells due to the additional interference caused by void cells [35–37]. It was found in [38–44] that with conservative offloading strategy, only limited performance gain can be achieved from the

deployment of small-scale cells. However, aggressively offloading mobile users from macro BSs to smaller BSs such as WiFi access point (AP) can lead to a severe degradation of the network performance. For example, a WiFi AP with excellent signal strength may suffer from heavy load or have less effective bandwidth (channels), thus reducing the effective rate it can serve at [45]. It is quite clear that both of the aforementioned consequences is undesirable, which motivates engineers to seek an optimal offloading strategy.

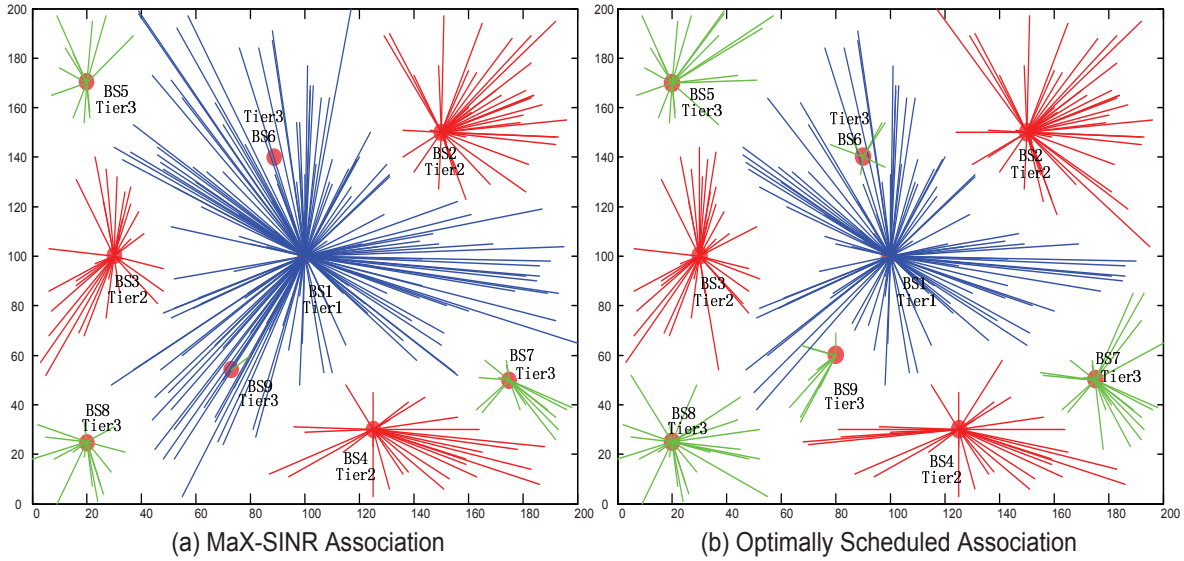


Figure 1.2: Illustration of the user-BS association scheme. (a) Max-SINR association. (b) Optimally scheduled association.

As a key component to realize the potential of capacity enhancement with the architecture of HetNets, load balancing has long been studied and attracted extensive attention. One direct approach for load balancing is to schedule each user-BS link in a centralized manner. By assuming no handover and without any selection criteria, each user could choose one BS to associate with freely. As demonstrated in Fig. 1.2, max-SINR association would not only overload the Tier-1 BS (macro BS) but also force small-scale BSs to serve very few users with some even being idle; By searching over all possible user-BS associations and finding the optimal one, the load pressure of Tier-1 BSs can be effectively balanced by its surrounding small-scale Tier-2 and Tier-3 BSs [46].

However, since the above general approach of the user-BS association subject to a resource would result in a coupled relationship between the users association and scheduling, the NP hard and combinatorial nature of this optimization problem was identified in [47–49]. In HetNets, pico and femto BSs are usually deployed with a much larger density than macro BSs. Users thus have more freedom to make the association choice. As the network scales up, finding an optimal solution to the combinatorial problem becomes untractable. Besides being computationally daunting, this approach is unlikely to lead to insight into the role of key parameters on system performance. Therefore, a few key mathematical approaches, i.e., relaxed optimization [46, 50], Markov decision process (MDP) [51–53], game theory [54, 55] have been applied.

Nevertheless, the above mentioned approaches, i.e. relaxed optimization, MDP and game theory, focus on strategic decision making of each individual user, and thus may have limitations to further characterize the relations between performance metrics and system parameters. Furthermore, these approaches are quite sensitive to the locations of users and BSs, indicating that the algorithms have to run over and over again in order to keep tracking of changes such as user mobility in networks. Hence, Cell Range Expansion (CRE) [34] was adopted to serve as an easy-to-implement technique for load balancing in HetNets. In early works regarding the user association problem, Cell Range Expansion was first proposed in code-division multiple access (CDMA) systems. This so-called "Cell Breathing" technique adopts a biased user-BS association scheme where each user assigns a biased value to the measured received power from pilot signals transmitted by available BSs, and associates with the BS that has the largest biased SINR [56, 57]. By doing so, uneven load conditions in macro cells can be equalized readily by the expansion or contraction of the cells coverage area. Similar approach was proposed in [58] where all the cells coordinate to adjust their coverage by using the Cell Breathing technique based on their load fluctuations. As the network architecture evolves into a heterogeneous form,

BSs are divided into multiple tiers according to their transmission powers. Therefore, BSs from different tiers have distinct cell sizes and load conditions, and thus they should be assigned with distinct SINR biasing factors [59].

Formally, if there are K candidate tiers available with which a user may associate, the index of the chosen tier is $k^* = \arg \max_{i=1,\ldots,K} B_i P_{R,i}$, where B_i is the biasing factor for Tier i and $P_{R,i}$ is the received power from Tier i. Conventionally, macro BS is usually denoted by Tier 1 and has a bias of 1 (0dB). For example, a 10dB bias of a small-scale BS indicates that a mobile user would associate with the small BS unless its received power is at least 10dB less than that of the macro BS. Biasing effectively expands the coverage area of small cells, so it is referred to as biased association scheme, which will be specified in the following. Fig. 1.3 demonstrates the technique of Cell Range Expansion of a 2-Tier HetNet.

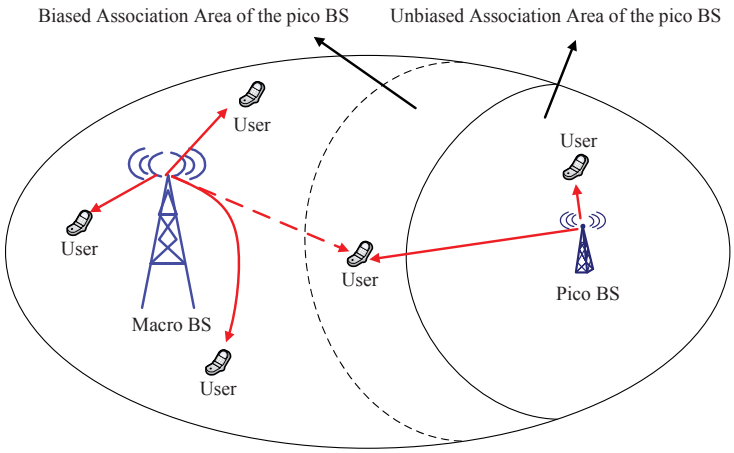


Figure 1.3: An example of a 2-Tier HetNet with Cell Range Expansion.

Analytical approaches for biasing and interference coordination were first studied in [60, 61]. However, the downlink rate was not investigated, which is one of the key metrics in evaluating the network performance. In [46], the authors formulated an optimization problem in HetNets exploiting a logarithmic utility function of users' long-term perceived rate to account for the proportional fairness [62], and obtained the optimal per-tier biasing

factor by a brute force search. It was shown that if the biasing factors are designed carefully, the Cumulative Distribution Function (CDF) of the overall user rate is pretty close to that achieved by solving the combinatorial association problem. However, this method depends on specific network realization and the optimal biasing factor was found empirically.

To find tractable expressions of key system performance metrics, stochastic geometry [63,64] was then introduced to model the locations of BSs and users as spatial random point process. There has been considerable achievement in the theory of HetNets whereby the locations of APs of each tier as well as the users are assumed to form a homogeneous Poisson point process (PPP). In [65], the authors considered a network topology of K independent tiers of Poisson point process (PPP) distributed BSs and derived an expression of SINR coverage, i.e., the probability that the SINR of a user exceeds a threshold value. Motivated by [65], optimal per-tier biasing factors was characterized by [66] by maximizing the rate coverage, i.e., the probability that a randomly chosen user can achieve a target rate. Similarly, [67] derived an explicit expression of the rate coverage by assuming resource partitioning, and numerically obtained the optimal biasing factor and the fraction of resource partitioning. By maximizing a network-wide proportional fair utility function based on the logarithm of the mean user rate, [68] analytically obtained the optimal biasing factors of each tier.

1.2.2 Review of Delay Optimization in HetNets

Even though the subject of user association in a HetNet has been well studied, most of the previous works did not take the QoS requirements into account explicitly. For a HetNet with QoS flows, optimizing the packet delay performance is even more relevant than maximizing the typically assumed proportional fair metric [69, 70]. Meanwhile, stringent delay requirement has been posed on the network nowadays due to the emergence of

new types of applications [1], such as latency-critical applications like command-and-control of drones, advanced manufacturing, and tactile Internet [71, 72]. In practice, with the proliferation of real-time multimedia applications, the packet delay is becoming an important quality-of-service (QoS) metric. For example, an end-to-end latency over 200 ms for real-time video media stream is generally considered to be unacceptable [73]. Therefore, a deeper understanding of the effect of key network parameters on packet transmission delay becomes essential to evaluate the overall network performance [74, 75].

Since previous studies [30, 38–44, 46–55, 63–65, 67, 68] assumed that the BSs always have packets to transmit, they present the network performance metrics of SINR and rate, and neglect the characterization of the network delay performance. However, there is an increasing interest in delay analysis for various types of networks. In particular, [76] investigated the delay performance of both resource separation and resource sharing strategies in terms of the provided QoS level of the internet. In [77], an adaptive scheme to schedule the delay-sensitive traffic in IEEE 802.11e Wireless Local Area Networks (WLAN) was proposed where the packets were queued based on their deadline to reach the destination. [78, 79] studied the local delay, i.e., the time it takes a node to successfully transmit a packet to its neighbor, in ad hoc Poisson networks with the consideration of node mobility. As for wireless networks, the packet blocking probability and the packet queuing delay were characterized in [80, 81] for a isolated traditional macro cell. In HetNets, the expected delay was analyzed in [82] by taking into account the delays in radio access and backhaul links. The optimal spectrum allocation pattern was obtained in [83, 84] by minimizing the average packet queuing delay. Similarly, Cheng *et al.* [85] derived a distributive stochastic learning algorithm to determine the optimal user scheduling and power control policy by minimizing the average network delay.

The most challenging part of characterizing delay performance is queuing analysis. The queuing performance of a single cell was evaluated in [80, 81] for the first time in CDMA

systems. By assuming constant interference over the entire cell, their work characterized the performance of only one independent queue. Nevertheless, such queuing analysis is not applicable to HetNets. In HetNets, as BSs of various types are quite close to each other due to a large deployment intensity, the queues of all co-channel BSs are spatially and temporally correlated, which is induced by interference and traffic/load patterns [86, 87]. In particular, each BS will only act as an interferer if it is in the busy state, leading to the coupled queue problem [88]. The analysis of coupled queues is a long-standing open problem, and even solving a special case of two interacting queues is challenging [89]. To solve the coupled queue problem, [83–85] considered fixed number and locations of BSs and modeled them as a n–dimension continuous time Markov chain (CTMC) based on the instantaneous channel state information (CSI) and queue state information. However, since CTMC can only deal with limited queues, the computational complexity becomes unbounded as the network scales up. Hence, stochastic geometry should be combined with queuing theory to decouple the queuing performance in HetNets, which will be elaborated in the next chapter.

1.2.3 Review of Energy Efficiency Optimization in HetNets

Besides the efforts to optimize network spectrum efficiency and delay, more and more intensive attention has been paid to improve the network energy efficiency since global energy consumption due to information and communication technologies is rising rapidly [90]. It is estimated in [91–95] that an active macro BS consumes 40 to 80 watts on transmission. By combining the power consumed for signal processing, computation, cooling, and radio frequency power amplification, the total power consumption can sums up to over 1000 watts. With the fruition of HetNets, deploying small-scale BSs with a huge density brings about a higher power consumption. According to [96], the current HetNet consumes approximately 60 billion kWh per year and is expected to double by

the year 2020.

Maximizing the energy efficiency has long been studied by previous literatures for cellular networks [97–101]. [97, 98] found the optimal BS deployment density by maximizing the ratio between BSs' achievable rate and the cellular network power consumption. Subject to SINR coverage and rate coverage, [99] first derived BSs' minimum transmission power, then obtained the optimal BS density to minimize the average network power consumption. Besides optimizing the BS density, [100, 101] focused on optimizing the operation mode (on/off status) of each BSs according to the load condition. In [100], the authors proposed a distributed on/off switching based algorithm in cellular networks to decide the minimum set of active BSs. By arguing that a cellular BS could operate in normal mode, sleep mode, or expansion mode, [101] proposed a scheme that determines which mode the BS should choose based on the load condition, such that the energy consumption is minimized.

As for HetNets, enhancing the network energy efficiency becomes more critical as the proliferation of small-scale BSs can cause a significant burden on the power consumption. On the other hand, since the small-scale BSs usually serve fewer users due to the limited association region, traffic fluctuation have a severer negative impact on the energy consumption in HetNets. In particular, the amount of user service requests can drop dramatically during non-peak traffic hours. The BSs are thus more likely to be idle during such periods, but still consume energy [102]. To reduce the total power consumption, [103] dynamically change the operating states (on and off) of the small-scale BSs, while keeping the macro BSs on to avoid any service failure outside active small cells. [104] proposed a scheme to determine the smallest set of BSs, which can be carried out periodically to adapt to aggregate traffic variations. The core idea of [103, 104] is to dynamically switch off a fraction of cells during periods of low activity load, which is quite similar to [100, 101].

However, as [103, 104] assumed fixed locations and number of BSs of all tiers, their

methods thus highly depend specific network realization. Therefore, to find tractable expressions of key system performance metrics, stochastic geometry was then adopted by [105–108] to model the irregular deployment of BSs as Poisson Point Process (PPP). In particular, [105,106] focused on a 2-tier HetNet and assumed that a transmission fails if the received SIR of a user is lower than a given threshold. Instead of directly minimizing the network power consumption, [105, 106] defined and maximized the network energy efficiency performance as the ratio between the network total power consumption and the network throughput, i.e., the average successfully transmitted bits per sec per Hz per unit area. Although the optimization problem is not convex, an iterative algorithm is proposed to obtain the optimal BS density of each tier. [107,108] also considered a HetNet consisting of 2 types of BSs, following independent PPP distributions. To avoid the coverage hole caused by BS sleeping, some BSs, called "coverage BSs", cannot be switched off. Under the network SINR coverage constraint, the authors in [107, 108] optimized the BS density in order to save energy. It is found that if the ratio of operation cost between micro and micro BSs are lower than a threshold, which is a function of transmission power and path loss, an optimal fraction of macro BSs should be obtained and switched off; otherwise, the strategy is the opposite, i.e., an optimal fraction of micro BSs should be calculated and powered off.

1.3 Challenges and Motivations

In this section, we will discuss the existing problems and challenges according to our literature review, based on which the motivations of this thesis is given.

How to balance the load pressure across tiers with queuing taken into account

Although plenty of efforts [46, 50–55, 60, 61, 65–68] have been made to strike a balanced load across tiers for a higher spectrum efficiency in HetNets, they assumed that the BSs are transmitting packets all the time without queuing considered inside BSs. Therefore, they all focused on the performance metrics such as rate and SINR. In practice, one BS can vary between busy and idle states over a small time scale due to the dynamic packet arrivals of its associated users, under which case one BS would not interfere with others unless it is busy transmitting packets to its user. As more small-scale cells are incorporated into macro cells to form a HetNet, fewer users are served by BSs. Therefore, it is of high probability that one BS can be idle during some time period such that the delay performance cannot be neglected. How to balance the BS load pressure across tiers in terms of the network delay performance by considering queuing thus becomes an interesting issue that deserves much attention.

As mentioned before, the most challenging part in characterizing the delay is to solve the coupled queue problem. Current queuing analysis in HetNets are all based on continuous-time Markov chain (CTMC) [83–85]. However, CTMC can only deal with specific network realization, no tractable expressions of the network performance can be derived. In addition, as the network scales up, the state space of the Markov process may become huge, and the analysis would become intractable. Hence, this motivates us to deal with the coupled queue problem with the tool of stochastic geometry to account for the random BS deployment, and derive the delay performance metrics analytically such that some insights can be gained for load balancing with the consideration of queuing.

How to maximize network energy efficiency with queuing taken into account

Since [97–101, 103–108] all considered fixed power consumption of BSs without taking queuing into account, the only system parameter that determines the network power

consumption is the BS deployment. The overall network power consumption thus linearly increases as the number/density of BSs increases. As a result, the energy efficiency optimization problem in [97–101, 103–108] falls into two categories:

- Minimize the network power/energy consumption to optimally switch off a fraction of BSs under the constraint of network rate and SINR performance;
- Maximize the ratio between the network power consumption and the network throughput to find the optimal BS deployment density.

By assuming dynamic traffic arrivals and queues inside BSs, nevertheless, other system parameters such as bandwidth allocation could account for the network power consumption performance as one BS consumes less energy in the idle state than it does in the busy state [95, 109]. Furthermore, as BSs are more likely to be idle with a larger deployment intensity, increasing the BS density would not necessarily deteriorate the energy efficiency. Therefore, how system parameters could impact on the network energy efficiency under the assumption of queuing needs to be reconsidered carefully.

How to optimize user-BS association with non-uniform user distribution

As mentioned before, most of the previous studies on load balancing [46, 50–55, 60, 61, 65–68] assumed a uniform user distribution. Therefore, these studies all adopt a per-tier biasing, i.e., BSs of a tier use the same biasing factor, as the traffic load of a tier is approximately the same. However, users might not be evenly distributed. In particular, a cluster of users might appear within the association region of a cell. For instance, the association region of a cell can be one hall or one room where people attend a lecture or enjoy a concert and thus form a cluster. In such case, the resource of a cell will be equally shared by more users than usual, which significantly lowers users' perceived rate. Hence, the tuning of the biasing factor in a per-tier fashion would not relieve the traffic pressure in the overloaded areas, and a per-station biased scheme is thus preferable. The

situation is quite intuitive: all of us must have experienced large drops in throughput due to congestion in crowded events, irrespective of signal quality, i.e., I have five bars but I cannot open a simple webpage. To enhance the spectrum efficiency of this overloaded cell as well as the entire network, the research question lies in "how to optimally tune the biasing factor of the overloaded BS according to load condition?".

1.4 Thesis Contributions and Structure

1.4.1 Thesis Contributions

To address the existing problems and open challenges elaborated in Section 1.3, this thesis aims to improve the network spectrum efficiency, delay performance and the network spectrum efficiency by focusing on a more practical scenario with both queuing in the BS and non-uniform user distribution. The key contributions of this thesis are summarized as follows.

1) Characterization of the BS average traffic intensity. To account for the irregular deployment of the BSs, stochastic geometry is adopted such that BSs of each tier are modeled as a homogeneous PPP. In contrast to previous studies where one BS is transmitting packets all the time, we consider that the packet requests from the users form a queue in their associated BSs. The traffic intensity of one BS thus varies with the aggregate packet requests of all its associated users. To decouple the queuing behavior of BSs, we resort to the approximation of replacing each BS's individual traffic intensity with the average traffic intensity of its tier. The spatial distribution of BSs in the busy state can thus be approximately characterized by a thinned-PPP model. The SIR coverage of each tier is then obtained, based on which the average traffic intensity of each tier is further obtained. It is further shown that when the spectrum resources is fully reused over the network, the average traffic intensity of each tier can be determined by a set of fixed-point

equations; With spectrum partitioning across tiers, an explicit expression of the average traffic intensity of each tier can be derived.

2) Delay-optimal biased user association in HetNets. Based on the characterization of the average traffic intensity, an optimization problem is formulated to minimize the lower bound of the network mean queuing delay by tuning the biasing factor of each tier, which is shown to be a convex problem. When the mean packet arrival rate of each user is small, a closed-form solution is derived. The simulation results demonstrate that the network queuing performance can be significantly improved by properly tuning the biasing factor. It is further shown that the network mean queuing delay might be improved at the cost of a deterioration of the network signal-to-interference ratio (SIR) coverage, which indicates a performance tradeoff between real-time and non-real-time traffic in HetNets.

3) Queue-Aware Optimal Bandwidth Allocation in HetNets. To properly allocate the spectrum resources to BSs of each tier in HetNets with the consideration of queuing, optimization problems to minimize the network average power consumption and to maximize the network SIR coverage are formulated, which are shown to be convex and concave with respect to bandwidth allocation, respectively. When the mean packet arrival rate of each user is small, closed-form solutions to the optimization problems are obtained. Simulation results of a 2-tier HetNet demonstrate that the network average power consumption and the SIR coverage can be significantly improved by the optimal spectrum allocation. A tradeoff between energy efficiency and SIR coverage is further revealed, which provides insights regarding the interplay of these two performance metrics.

4) Queue-aware energy efficient base station density optimization in HetNets. By using the approximation that BSs of a tier have the same SIR coverage, the cumulative distribution function (CDF) of the traffic intensity of each tier is obtained. On that basis, a minimization problem of the network average power consumption is studied by optimally tuning the activation ratio of micro BSs under the quality of serive (QoS)

constraints of the network mean queuing delay and the network SIR coverage. Numerical results demonstrate that if the idle power coefficient is below a certain threshold, the optimal activation ratio should equal the one to minimize the network average power consumption. Otherwise, the optimal activation ratio should be obtained according to the QoS constraints. It is further shown that universal frequency reuse (UFR) outperforms spectrum partitioning (SP) in terms of both energy efficiency and SIR coverage.

5) Optimal biased association scheme with non-uniform user distribution in HetNets. A practical scenario is studied where one cell is overloaded due to the cluster of users. In this case, the adjustment of the per-tier biasing factor becomes unreasonable, and thus we propose to adjust the biasing factor of the overloaded cell to offload the traffic to its surrounding cells. By maximizing the mean user utility in the area of this overloaded cell and its neighboring cells, the optimal biasing factor can be obtained. It is found that in the scenario where the overloaded cell is fully surrounded by a macro cell, the optimal biasing factor logarithmically decreases with the user's intensity of the overloaded cell. Numerical results demonstrate that by using the optimal biasing factor of the overloaded cell in the considered scenario, both the mean user rate in the overloaded cell and the overall mean user rate can be improved compared to the previous biased scheme without the adjustment of the overloaded cell in the literature. The analysis provides guidance on the optimal tuning of the biasing factor of an overloaded cell and, is a step forward towards the goal of the adjustment of the biasing factor in a per-station fashion under non-uniform spatial user distribution.

1.4.2 Thesis Structure

The rest of this thesis is organized as follows. Queuing analysis for both universal frequency reuse and orthogonal spectrum partitioning is presented in Chapter 2. An optimal biased association scheme to minimize a lower bound of the network mean queuing delay

with queuing taken into account is studied Chapter 3. A queue-aware optimal bandwidth allocation across tiers to maximize the network SIR coverage and energy efficiency is examined in Chapter 4. A queue-aware energy efficient BS density optimization problem under the QoS constraints of network mean queuing delay and network SIR coverage is formulated and solved in Chapter 5. An optimal biased association scheme to optimally offload the users from the overloaded cell with non-uniform user distribution is proposed and studied in Chapter 6. Conclusions and future works are given in Chapter 7.

Chapter 2

Queuing Analysis

Throughout this thesis, we mainly focus on the scenario that one BS could either be busy or idle and the interference only comes from the BSs in the busy state, which has been mentioned in Chapter 1. As a result, the characterization of the network delay performance, the network spectrum efficiency as well as the network energy efficiency in this thesis are closely related to the queue status of each BS, which lays the foundation for the network performance characterization and optimization in the following chapters.

In particular, Section 2.1 first identifies the queuing model and the coupled nature of the queues. The mathematical approaches to decouple the queuing behavior of the BSs, i.e., stochastic geometry and independent thinning, are introduced in Section 2.2. The analytical results of the average traffic intensity of each tier for the cases of orthogonal spectrum partitioning and universal frequency reuse are derived in Section 2.3, which is then verified by a spatial-temporal simulation of a 2-Tier HetNet.

2.1 Coupled Queue Problem

2.1.1 Queuing Model

Consider a K-tier heterogeneous network where BSs in the kth tier are denoted by the set $\Phi_k = \{\mathrm{BS}_{k,1}, \mathrm{BS}_{k,2}, ..., \mathrm{BS}_{k,N_k}\}$, $k \in \{1, ..., K\}$, where N_k is the total number of the Tier-k BSs. The mobile users form another set $\Phi_{\mathrm{u}} = \{\mathrm{UE}_1, \mathrm{UE}_2, ..., \mathrm{UE}_{N_{\mathrm{u}}}\}$, where N_{u} is the total number of the users. For a random user $\mathrm{UE_i}$ located at the origin, the instantaneous received power from a typical $\mathrm{BS_{k,j}}$ in the k^{th} tier is given by

$$P_{R,\{k,j\}} = P_k g_{k,j} x_{k,j}^{-\alpha_k}, \tag{2.1}$$

where $x_{k,j}$ denotes the distance between $\mathrm{UE_i}$ and $\mathrm{BS_{k,j}}$; P_k is the transmission power of a BS in the k^{th} tier; $g_{k,j}$ denotes the small-scale fading coefficient, which follows an i.i.d. exponential distribution of unit mean; and α_k is the path-loss coefficient, which is assumed to be identical across different tiers, i.e., $\alpha_k = \alpha$, $\forall k$. We assume in this chapter that each user associates with the BS with the largest average reference signal receiving power (RSRP)[1]. For the resource allocation, we assume that BSs of the same tier share the spectrum with a bandwidth of W_k, $k \in \{1, ..., K\}$. Denote the total bandwidth as W. With spectrum partitioning (SP) across tiers, we then have $\sum_{k=1}^{K} W_k = W$. With universal frequency reuse (UFR) over the network, we have $W_k \equiv W$ for each $k \in \{1, ..., K\}$.

For each user in the network, assume that its packet requests follow an independent Poisson process with a mean arrival rate γ, and the packet length is exponentially distributed with mean L. The incoming packets for all users form a queue in the associated BS, and the BS will transmit these packets in a first-in-first-serve (FIFS) fashion. To avoid users in poor channel conditions occupying the BS, we consider a fixed rate modulation and coding format. In particular, a BS will serve a user only when its instantaneous SIR

[1]Note that the queuing analysis based on the largest RSRP in this chapter can be applied to a biased association scheme, which will be shown in Chapter 3.

exceeds a threshold τ, and will drop its packet request otherwise. Note that due to a high BS deployment intensity, the background noise is dominated by the interference, and is therefore ignored in this thesis. According to Shannon's formula, the service rate for each user that is associated to a Tier-k BS can be obtained as

$$\mu_k = \frac{W_k}{L}\log_2\left(1+\tau\right). \tag{2.2}$$

For a randomly selected Tier-k BS, $\mathrm{BS}_{k,i}$, its traffic intensity, $\rho_{k,i}$, can be obtained as

$$\rho_{k,i} = \frac{\gamma_{k,i}}{\mu_k}, \tag{2.3}$$

where $\gamma_{k,i}$ is the mean aggregate packet arrival rate of all its associated users. Note that $\rho_{k,i}$ can also be interpreted as the busy probability or the utilization of $\mathrm{BS}_{k,i}$ when $\rho_{k,i} \leq 1$. Since $\mathrm{BS}_{k,i}$ delivers a packet only when the SIR exceeds a certain threshold τ, its mean aggregate packet arrival rate can be obtained as

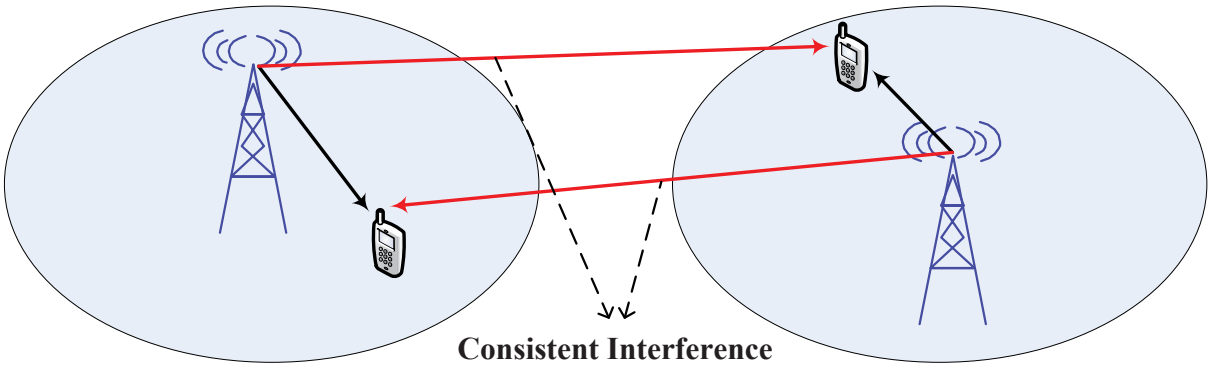
$$\gamma_{k,i} = \gamma N_{k,i}\mathrm{Pr}\left[\mathrm{SIR_{k,i}} > \tau\right], \tag{2.4}$$

and where $N_{k,i}$ is the number of users that are associated to $\mathrm{BS}_{k,i}$ and $\mathrm{Pr}\left[\mathrm{SIR_{k,i}} > \tau\right]$ denotes the SIR coverage of $\mathrm{BS}_{k,i}$, i.e., the probability that the SIR of a random user associated to $\mathrm{BS}_{k,i}$ is larger than the threshold τ. By substituting (2.4) into (2.3), the traffic intensity $\rho_{k,i}$ can be further written as

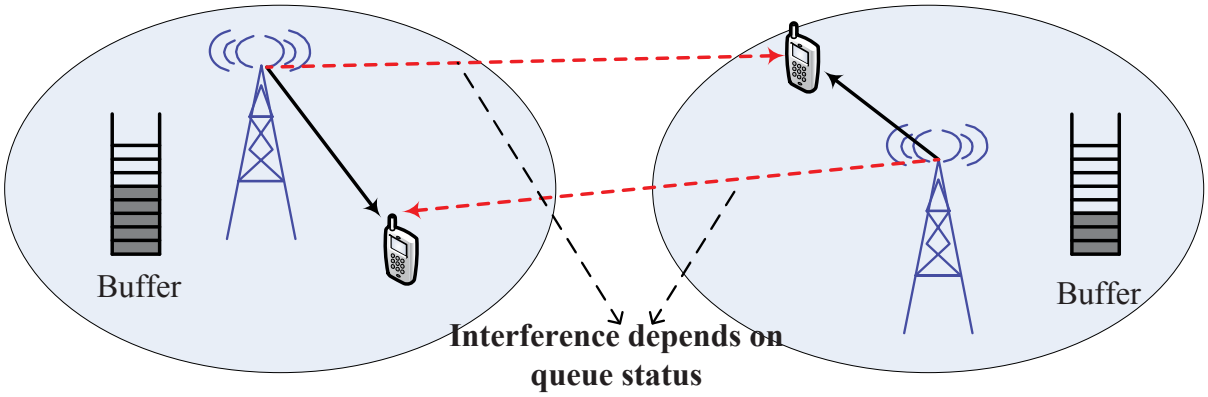
$$\rho_{k,i} = \frac{\gamma N_{k,i}\mathrm{Pr}\left[\mathrm{SIR_{k,i}} > \tau\right]}{\mu_k}, \tag{2.5}$$

As the BS will be always be active and the queue in the BS will be unstable if $\rho_{k,i} > 1$, we focus on the condition $\rho_{k,i} \leq 1$ in this thesis. In this case, $\rho_{k,i}$ equals the busy probability of the BS. Due to a varied association region, each BS has a different mean aggregate packet arrival rate, and the traffic intensity $\rho_{k,i}$ varies across each BS. In addition, as the experienced interference of one typical BS comes from all other co-channel BSs that are currently transmitting, there exists a spatial-temporal correlation of the queues of the BSs, which will be demonstrated in the following.

2.1.2 Spatial-Temporal Correlation of the Queues



(a)



(b)

Figure 2.1: Interference pattern between the neighboring BSs. (a) Without queuing considered (b) With queuing considered.

To study the spatial-temporal correlation of the coupled queues, let us first consider two neighboring BSs which shares the spectrum resources, which is illustrated in Fig. 2.1. As a comparison, Fig. 2.1(a) demonstrates the interference pattern under the assumption that BSs always have packets to transmit. Hence, there exists consistent interference from the neighboring BS to a typical user. With queuing considered as shown in Fig. 2.1, nevertheless, each BS will transmit the packets in a FIFO fashion, under which case the interference depends on the queue status of the BSs. Fig. 2.1 demonstrates a simple case

of the coupled queue problem. Intuitively, when the first BS transmits, it increases the interference experienced by the second BS and hence reduces its data rate; As a result, the second BS now takes longer to transmit same amount of data than it would have taken if the first BS was not transmitting. Hence, the queues are correlated and traffic intensities of both BSs are coupled.

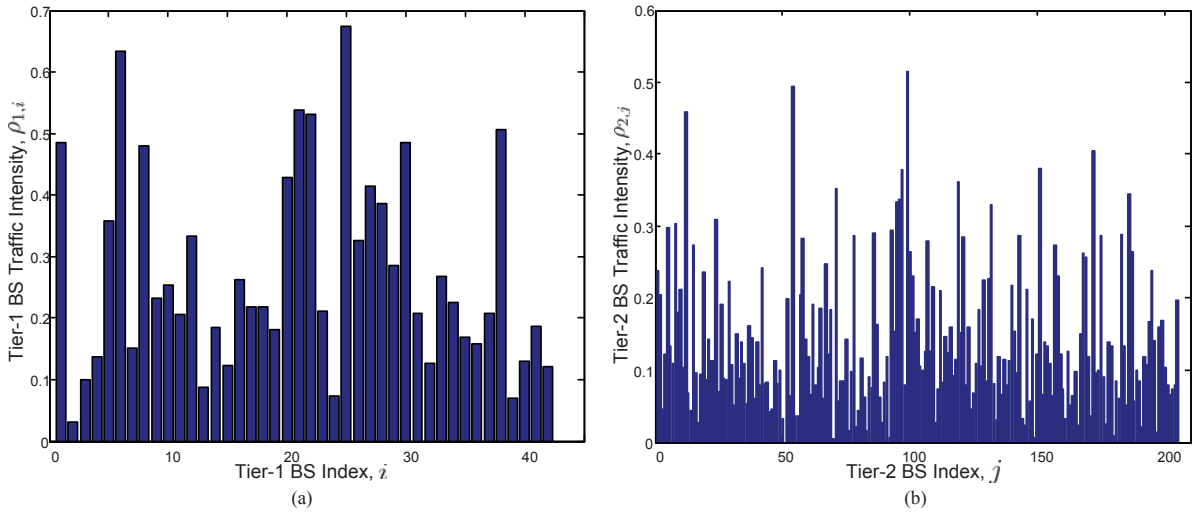


Figure 2.2: Traffic intensity of each BS in one simulation run. (a) Tier-1 BS traffic intensity $\rho_{1,i}$, $i \in \{1, ..., N_1\}$. (b) Tier-2 BS traffic intensity $\rho_{2,j}$, $j \in \{1, ..., N_2\}$.

With fixed locations of the BSs and the users in one specific K-Tier heterogeneous network realization and spectrum partitioning across tiers, the BS traffic intensity $\rho_{k,i}$ is a function of the traffic intensities of all other BSs in the same tier, i.e.,

$$\rho_{k,i} = f\left(\rho_{k,1}, ..., \rho_{k,i-1}, \rho_{k,i+1}, ..., \rho_{k,N_k}\right), \tag{2.6}$$

where $i \in \{1, ..., N_k\}$. With universal frequency reuse, $\rho_{k,i}$ is a function of the traffic intensities of all other BSs over the network, i.e.,

$$\rho_{k,i} = f\left(\rho_{1,1}, ..., \rho_{k-1,N_{k-1}}, \rho_{k,1}, ..., \rho_{k,i-1}, \rho_{k,i+1}, ..., \rho_{k,N_k}, \rho_{k+1,1}, ..., \rho_{K,N_K}\right), \tag{2.7}$$

where $i \in \{1, ..., N_k\}$. Fig. 2.2 illustrates the simulation results of the traffic intensity $\rho_{k,i}$ of each Tier-k BS in one fixed network realization. The total number of Tier-1 and

Tier-2 BSs are $N_1 = 42$ and $N_2 = 205$, respectively, and the total number of mobile users is $N_{\mathrm{u}} = 4000$. Each BS and user is randomly located in a square area of $4{*}10^6\mathrm{m}^2$. The transmission powers of each BS in the two tiers are given by $P_1 = 20\mathrm{W}$ and $P_2 = 6\mathrm{W}$, respectively. Each user then associates to their BSs by the largest average reference signal receiving power. For demonstration, the bandwidth allocation of the two tiers is given by $W_1 = W_2 = 6\mathrm{MHz}$. Other system parameters are set to be $\alpha = 4$, $\tau = 1$, $\gamma = 60\mathrm{Packets/s}$, $L = 0.001\mathrm{Mb}$. It can be observed from Fig. 2.2 that the traffic intensity $\rho_{k,i}$ of each BS in the k^{th} tier varies due to different cell sizes and the spatial-temporal correlations with other BSs. For a typical $\mathrm{BS}_{k,i}$ in the k^{th} tier, the only way to obtain its traffic intensity $\rho_{k,i}$ is to solve the set of equations (2.6). When N_k becomes large, solving (2.6) directly is intractable. To analyze the queuing performance of the BSs, mathematical approaches and approximations such as stochastic geometry should be adopted, which will be shown in the following.

2.2 Methodology to Decouple the Correlation

2.2.1 Stochastic Geometry

As mentioned in Section 2.1.2, the basic challenge of solving (2.6) or (2.7) is the huge computational complexity, which grows unbounded as the network scales up. Besides, since the solution of (2.6) or (2.7) highly depends on specific network realization, no insight could be given on the impact of key system-level parameters like transmission power, BS deployment density, and bandwidth allocation on the design of load balancing. To overcome those disadvantages, we adopt stochastic geometry [63, 64] in this thesis to account for the irregular locations of BSs and users in HetNets. Readers can refer to Table 2.1 for quick access to the notations used in this thesis.

Stochastic geometry is a probabilistic analytical approach, where the network config-

Table 2.1: Major Notation Summary

Notation	**Description**
Φ_k; Φ_{u}	PPP of active Tier-k BSs; PPP of users
λ_k; λ_{u}	Deployment density of Tier-k BSs; Density of users
α_k	Path-loss coefficient of Tier k
W	Total Bandwidth
γ	Mean packet arrival rate of each user
L	Mean packet length
τ	SIR threshold
$\rho_{k,i}$; $\bar{\rho}_k$	Traffic intensity of i-th BS in Tier k; Average traffic intensity of Tier k
S_k; S; $\hat{S}$	SIR coverage of Tier k; Network SIR coverage; Threshold of network SIR coverage
D_k; D; $\hat{D}$	Mean queuing delay of Tier k; Network mean queuing delay; Threshold of network mean queuing delay
P_k; $P_{k,s}$; Δ_k	Transmission power of Tier-k BSs; Additional power consumption of Tier-k BSs; Power Consumption Coefficient
$P_{k,\mathrm{T}}$; $P_{k,\mathrm{I}}$; $\bar{P}_k$; P	Power consumption of an active Tier-k BS in busy state; Power consumption of an active Tier-k BS in idle state; Average power consumption of an active Tier-k BS; Network average power consumption per area
η_k	Idle power coefficient of Tier-k BSs

uration is assumed random and following a certain distribution. In particular, the set of BSs in tier Φ_k, $k \in \{1, ..., K\}$, is assumed to follow an independent 2–D homogeneous Poisson Point Process (PPP) with a given intensity λ_k [65]. Similarly, the set of users Φ_{u} can be modeled as another independent PPP with a given intensity λ_{u}. This has the advantage of avoiding specific information on the network topology. In particular, for a given area, the total numbers of the BSs of each tier N_k as well that of users N_{u} are random, and the their locations are also randomly and uniformly distributed in such considered area. The only system parameter that determines the HetNet is the deployment intensity λ_k of the BSs in each tier and the intensity of the users λ_{u}. Fig. 2.3 illustrates the PPP characterization of 3-tier HetNet. Note that red dots are referred to as macro BSs, green triangles are referred to as pico BSs, and dark squares are referred to as femto BSs. It has been shown in [110, 111] that modeling BSs of each tier as independent PPP distributions achieves a high accuracy towards real wireless networks.

Due to the random topology by adopting stochastic geometry as a tool, the focus on the exact traffic intensity $\rho_{k,i}$ of each BS in the set Φ_k is then converted into the average traffic intensity performance, i.e., from $\{\rho_{k,i}\}_{i\in\Phi_k}$ to $\underset{i\in\Phi_k}{E}[\rho_{k,i}]$ for each $k \in \{1, ..., K\}$. By combining (2.5), the average traffic intensity of Tier k can be is given by

$$\bar{\rho}_k = E[\rho_{k,i}] = E\left[\frac{\gamma N_{k,i} \Pr[\mathrm{SIR_{k,i}} > \tau]}{\mu_k}\right]$$

$$= \frac{\gamma}{\mu_k} E[N_{k,i}] \, E[\Pr[\mathrm{SIR_{k,i}} > \tau]] = \frac{\gamma L \bar{N}_k S_k}{W_k \log_2(1+\tau)}, \tag{2.8}$$

where $\bar{N}_k = E[N_{k_i}]$ denotes the average number of users associated with a Tier-k BS and $S_k = E[\Pr[\mathrm{SIR_{k,i}} > \tau]]$ denotes the SIR coverage of all Tier-k BSs, i.e., the probability that the SIR of a typical user associated with a Tier-k BS exceeds the threshold τ. As the average traffic intensity, $\bar{\rho}_k$, is determined by the average number of associated users, $\bar{N}_k$, and the SIR coverage, S_k, we will derive these two components in the following.

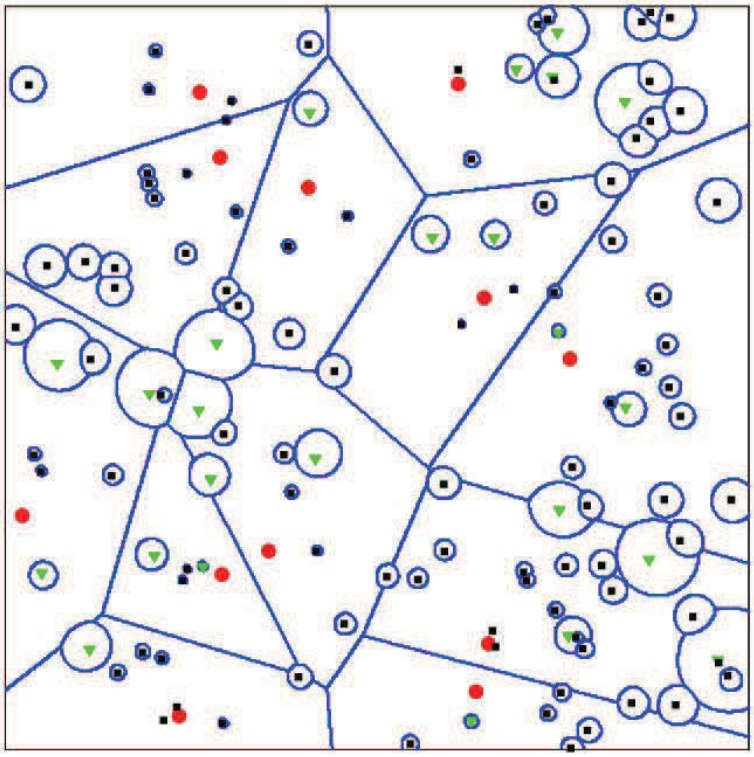


Figure 2.3: Close-up view of coverage regions in a 3-tier HetNet.

According to [59], the average number of users associated with a Tier-k BS, $\bar{N}_k$, has been obtained as

$$\bar{N}_k = \frac{\lambda_{\mathrm{u}} A_k}{\lambda_k}, \tag{2.9}$$

where A_k denotes the probability for a typical user to be associated with a Tier-k BS. Note that the association probability A_k has been derived in [59] as

$$A_k = \frac{\lambda_k (P_k)^{2/\alpha}}{\sum_{j=1}^{K} \lambda_j (P_j)^{2/\alpha}} = \frac{1}{\sum_{j=1}^{K} \widetilde{\lambda}_j \left(\widetilde{P}_j\right)^{2/\alpha}}, \tag{2.10}$$

where $\widetilde{\lambda}_j = \lambda_j/\lambda_k$ and $\widetilde{P}_j = P_j/P_k$ denote the normalized intensity and the normalized transmission power, respectively, conditioned on Tier k being a serving tier.

To characterize of the Tier-k SIR coverage S_k, some approximation approach, i.e., independent thinning, should be adopted in this thesis, which will be introduced in the next section.

2.2.2 Independent Thinning

Recall that BSs of Tier k form a PPP Φ_k with an intensity of λ_k. Moreover, for a randomly selected $\text{BS}_{k,i}$ where $i \in \Phi_k$, the traffic intensity $\rho_{k,i}$ can be interpreted as its busy probability when $\rho_{k,i} \leq 1$. The set of Tier-k BSs being busy, therefore, forms a thinned point process $\Phi'_k \subseteq \Phi_k$ by including $\text{BS}_{k,i} \in \Phi_k$ with the probability $\rho_{k,i}$ [112]. Since the traffic intensity of one BS is different from each other, the thinned point process Φ'_k is non-homogeneous. For analytical tractability, we adopt the independent thinning approach to approximately regard Φ'_k as an independently thinned homogeneous Poisson point process, which is thinned by the average traffic intensity $\bar{\rho}_k$ of this tier. From this approach, the intensity of the approximated homogeneous PPP Φ'_k is given by

$$\lambda'_k = \bar{\rho}_k \lambda_k. \tag{2.11}$$

It will be demonstrated in Section 2.3.3 that the independent thinning approach achieves a good approximation in deriving the average traffic intensity $\bar{\rho}_k$.

2.3 Average Traffic Intensity

Based on the mathematical approaches of stochastic geometry and independent thinning, we derive the expressions of the average traffic intensity $\bar{\rho}_k$ for both the cases of orthogonal spectrum partitioning and universal frequency reuse in this section. Then we perform a spatio-temporal simulation to justify our derived analytical results.

2.3.1 Orthogonal Spectrum Partitioning

For a typical user that is associated with a Tier-k BS, the interference all comes from busy BSs of the same tier. According to (2.1), the SIR of this typical user can then be

written as

$$\mathrm{SIR}_k = \frac{P_k g_{x_{k,0}} x_{k,0}^{-\alpha}}{\sum\limits_{j \in \Phi_k' \backslash \mathrm{BS}_{k,0}} P_k g_{k,j} x_{k,j}^{-\alpha}}, \tag{2.12}$$

where $x_{k,0}$ and $x_{k,j}$ denote the distance from the typical user to the associated BS $\mathrm{BS}_{k,0}$ and the jth interfering Tier-k BS, respectively; $g_{k,0}$ and $g_{k,j}$ denote the small-scale fading coefficient of $\mathrm{BS}_{k,0}$ and the jth interfering Tier-k BS, respectively. In (2.12), $\mathrm{BS}_{k,0}$ and $\Phi_k' \backslash \mathrm{BS}_{k,0}$ denote the associated Tier-k BS of this typical user and the set of interfering Tier-k BSs, respectively. Note that as spectrum partitioning is assumed across tiers, there is no inter-tier interference, and the interfering sources consist of all the busy Tier-k BSs besides the associated $\mathrm{BS}_{k,0}$. The following lemma presents the SIR coverage of a Tier-k BS.

Lemma 2.1. *If spectrum partitioning across tiers is adopted, the SIR coverage of a Tier-k BS can be written as*

$$S_k = \frac{1}{A_k \bar{\rho}_k Z(\tau, \alpha) + 1}, \tag{2.13}$$

where $Z(\tau, \alpha) = \tau^{\frac{2}{\alpha}} \int_{(1/\tau)^{\frac{2}{\alpha}}}^{\infty} \frac{du}{1+u^{\frac{\alpha}{2}}}$.

Proof. The probability density function (PDF) of the distance $x_{k,0}$ between a typical user and its serving Tier-k BS has been obtained in [59] as

$$f_{x_{k,0}} = \frac{2\pi\lambda_k}{A_k} x_{k,0} \exp\left(-\pi x_{k,0}^2 \frac{\lambda_k}{A_k}\right). \tag{2.14}$$

Using (2.14), the SIR coverage of a Tier-k BS can be obtained as

$$\begin{aligned} S_k &= \int_0^{\infty} S_k(x_{k,0}) f_{x_{k,0}} dx_{k,0} \\ &= \int_0^{\infty} S_k(x_{k,0}) \frac{2\pi\lambda_k}{A_k} x_{k,0} \exp\left(-\pi x_{k,0}^2 \frac{\lambda_k}{A_k}\right) dx_{k,0} \end{aligned} \tag{2.15}$$

where $S_k(x_{k,0})$ is the SIR coverage of Tier k conditioned on the distance between the typical user and the serving tier-k BS being $x_{k,0}$.

By denoting I as the set of interfering Tier-k BSs, i.e., $I = \Phi'_k \backslash \mathrm{BS}_{k,0}$, the conditional SIR coverage of Tier k can be written as

$$S_k\left(x_{k,0}\right) = \Pr\left[\frac{P_k g_{k,0} x_{k,0}^{-\alpha}}{\sum\limits_{j\in I} P_k g_{k,j} x_{k,j}^{-\alpha}} > \tau \middle| x_{k,0}\right]$$

$$\begin{aligned}
&\overset{(a)}{=} E\left[\exp\left(-\tau \sum_{j\in I} P_k g_{k,j} x_{k,j}^{-\alpha} P_k^{-1} x_{k,0}^{\alpha}\right) \middle| x_{k,0}\right] \\
&= E_{\Phi'_k, g_{k,j}}\left[\prod_{j\in I} \exp\left(-\tau x_{k,0}^{\alpha} g_{k,j} x_{k,j}^{-\alpha}\right) \middle| x_{k,0}\right] \\
&\overset{(b)}{=} \exp\left\{-\int_{\mathbb{R}^2} \left[1 - E_{g_{k,j}}\left[\exp\left(-\tau x_{k,0}^{\alpha} g_{k,j} x_{k,j}^{-\alpha}\right)\right]\right] \times \lambda'_k dj\right\}
\end{aligned} \tag{2.16}$$

according to (2.12), where (a) follows from the fact that $g_{k,0}$ is an exponential random variable with unit mean, and (b) follows from the probability generating functional (PGFL) of Φ'_k [112] due to the independency between Φ'_k and $g_{k,j}$. According to (2.11), (2.16) can be further written as

$$\begin{aligned}
&\exp\left\{-\lambda_k \bar{\rho}_k \int_{\mathbb{R}^2} \left[1 - E_{g_{k,j}}\left[\exp\left(-\tau x_{k,0}^{\alpha} g_{k,j} x_{k,j}^{-\alpha}\right)\right]\right] dj\right\} \\
&= \exp\left\{-2\pi\lambda_k \bar{\rho}_k \int_{x_{k,0}}^{\infty} \left[1 - E_{g_{k,j}}\left[\exp\left(-\tau x_{k,0}^{\alpha} g_{k,j} x_{k,j}^{-\alpha}\right)\right]\right] \times x_{k,j} dx_{k,j}\right\} \\
&\overset{(a)}{=} \exp\left[-2\pi\lambda_k \bar{\rho}_k \int_{x_{k,0}}^{\infty} \left(1 - \frac{1}{1 + x_{k,0}^{-\alpha} \tau^{-1} x_{k,j}^{\alpha}}\right) x_{k,j} dx_{k,j}\right] \\
&= \exp\left[-\pi \bar{\rho}_k \lambda_k x_{k,0}^2 Z\left(\tau, \alpha\right)\right],
\end{aligned} \tag{2.17}$$

where

$$Z\left(\tau, \alpha\right) = \tau^{2/\alpha} \int_{(1/\tau)^{2/\alpha}}^{\infty} \frac{du}{1 + u^{\alpha/2}}. \tag{2.18}$$

Note that (a) follows from the exponential distribution of $g_{k,i}$ with unit mean. Finally,

by combining (2.15) and (2.17), (2.13) can be obtained as

$$\begin{aligned} S_k &= \int_0^{\infty} \exp\left[-\pi\bar{\rho}_k\lambda_k x_{k,0}^2 Z(\tau,\alpha)\right] \cdot \frac{2\pi\lambda_k}{A_k} x_{k,0} \exp\left(-\pi x_{k,0}^2 \frac{\lambda_k}{A_k}\right) dx_{k,0} \\ &= \frac{1}{A_k\rho_k Z(\tau,\alpha)+1}. \end{aligned} \tag{2.19}$$

□

According to Lemma 1, the outage probability of Tier k, $O_k = 1 - S_k$, can be written as $O_k = \frac{A_k\rho_k Z(\tau,\alpha,1)}{A_k\rho_k Z(\tau,\alpha,1)+1}$. If Tier-$k$ BSs are always busy, i.e., $\bar{\rho}_k = 1$, the outage probability O_k reduces to the results in [59].

By combining (2.8), (2.9), and (2.10), the average traffic intensity $\bar{\rho}_k$ of Tier-k BSs can be explicitly derived as

$$\bar{\rho}_k = \frac{-\lambda_k R_k + \left[(\lambda_k R_k)^2 + 4\gamma\lambda_{\mathrm{u}}\lambda_k A_k^2 R_k L Z\right]^{\frac{1}{2}}}{2A_k\lambda_k Z R_k} \tag{2.20}$$

where Z denotes $Z(\tau,\alpha)$ for simplicity and $R_k = W_k \log_2(1+\tau)$ is the transmission rate of a Tier-k BS.

2.3.2 Universal Frequency Reuse

Different from spectrum partitioning, all the other active BSs in the busy state are the interfering sources under the assumption of universal frequency reuse. By denoting Φ'_j as the set of active Tier-j BSs in the busy state, the SIR of one typical user associated to a Tier-k BS can be written as

$$\mathrm{SIR}_{k,0} = \frac{P_k g_{x_{k,0}} x_{k,0}^{-\alpha}}{\sum_{j=1}^{K} \sum_{i\in\Phi'_j\setminus \mathrm{BS}_{k,0}} P_j g_{j,i} x_{j,i}^{-\alpha}}, \tag{2.21}$$

where $\mathrm{BS}_{k,0}$ and $\Phi'_j\setminus\mathrm{BS}_{k,0}$ denote the associated Tier-k BS and the interfering Tier-j BSs, respectively. $x_{k,0}$ and $x_{j,i}$ denote the distances from this typical user to its associated

$\mathrm{BS}_{k,0}$ and the ith interfering BS in the jth tier, respectively; $g_{k,0}$ and $g_{j,i}$ denote the small-scale fading coefficients of the channel to the associated Tier-k BS and the ith Tier-j interfering BS, respectively. The following lemma gives the SIR coverage of a Tier-k BS with universal frequency reuse.

Lemma 2.2. *If universal frequency reuse is adopted, the SIR coverage of a Tier-k BS is given by*

$$S_k = \frac{1}{A_k \sum_{j=1}^{K} \tilde{\lambda}_j \tilde{P}_j^{-\frac{2}{\alpha}} \bar{\rho}_j Z + 1}, \tag{2.22}$$

where $Z(\tau, \alpha) = \tau^{\frac{2}{\alpha}} \int_{(1/\tau)^{\frac{2}{\alpha}}}^{\infty} \frac{du}{1+u^{\frac{\alpha}{2}}}$ *and* $\tilde{P}_j$ *and* $\tilde{\lambda}_j$ *denote the normalized BS transmission power and the active BS density of Tier j conditioned on that of Tier k, respectively.*

Proof. By combining (2.21), the conditional SIR coverage of Tier k with a given $x_{k,0}$ can be written as

$$\begin{aligned}
&S_k(x_{k,0}) \\
&= E\left[\Pr\left[\mathrm{SIR}_{k,0} > \tau \middle| x_{k,0}\right]\right] \\
&= E\left[\Pr\left[\frac{P_k g_{k,0} x_{k,0}^{-\alpha}}{\sum_{j=1}^{2} \sum_{i \in \Phi'_j \setminus \mathrm{BS}_{k,0}} P_j g_{j,i} x_{j,i}^{-\alpha}} > \tau \middle| x_{k,0}\right]\right] \\
&\overset{(a)}{=} \prod_{j=1}^{K} E_{\Phi'_j, g_{j,i}}\left[\prod_{i \in \Phi'_j \setminus \mathrm{BS}_{k,0}} \exp\left(-\tau \tilde{P}_j g_{j,i} x_{j,i}^{-\alpha} x_{k,0}^{\alpha}\right)\right] \\
&\overset{(b)}{=} \prod_{j=1}^{K} E_{\Phi'_j}\left[\prod_{i \in \Phi'_j \setminus \mathrm{BS}_{k,0}} E_{g_{j,i}}\left[\exp\left(-\tau \tilde{P}_j g_{j,i} x_{j,i}^{-\alpha} x_{k,0}^{\alpha}\right)\right]\right] \\
&\overset{(c)}{=} \prod_{j=1}^{K} \exp\left\{-\int_{\mathbb{R}^2} \lambda'_k \left\{1 - E_{g_{j,i}}\left[\exp\left(-\tau P_j g_{j,i} x_{j,i}^{-\alpha} P_k^{-1} x_{k,0}^{\alpha}\right)\right]\right\} di\right\},
\end{aligned} \tag{2.23}$$

where $\tilde{P}_j = P_j / P_k$ is the normalized BS transmission power of Tier j conditioned on Tier k. Note that (a) follows from the fact that $g_{k,0}$ is an exponentially distributed random

variable with unit mean, (b) follows that Φ'_j and $g_{j,i}$ are independent random variables, and (c) follows the follows from the PGFL of Φ'_k with the intensity λ'_k.

Then, by substituting (2.11) into (2.23), $S_k\left(x_{k,0}\right)$ can be further written as

$$\begin{aligned} S_k\left(x_{k,0}\right) &= \prod_{j=1}^{K} \exp\left\{ -\lambda_k \rho_k \int_{x_{k,0}\tilde{P}_j^{\frac{1}{\alpha}}}^{\infty} \left[1 - E_{g_{j,i}}\left[\exp\left(-\tau \tilde{P}_j g_{j,i} x_{j,i}^{-\alpha} x_{k,0}^{\alpha} \right) \right] \right] x_{j,i} dx_{j,i} \right\} \\ &\overset{(a)}{=} \prod_{j=1}^{K} \exp\left\{ -2\pi\lambda_j\rho_j \times \int_{x_{k,0}\tilde{P}_j^{\frac{1}{\alpha}}}^{\infty} \left(1 - \frac{1}{1 + x_{k,0}^{-\alpha} \tilde{P}_j^{-1} \tau^{-1} x_{j,i}^{\alpha}} \right) x_{j,i} dx_{j,i} \right\} \\ &= \prod_{j=1}^{K} \exp\left\{ -\pi\lambda_j\rho_j \tilde{P}_j^{-\frac{2}{\alpha}} x_{k,0}^2 Z \right\}, \end{aligned} \tag{2.24}$$

where $Z = Z\left(\tau, \alpha\right)$ can be found in (2.18), and (a) follows from the fact that $g_{j,i}$ is an exponential random variable with unit mean. Finally, by combining (2.14), the SIR coverage of Tier k can be obtained as

$$S_k = \frac{1}{A_k \sum\limits_{j=1}^{2} \tilde{\lambda}_j \rho_j \tilde{P}_j^{-\frac{2}{\alpha}} Z + 1}. \tag{2.25}$$

□

By substituting (2.22) into (2.8), the average traffic intensity $\bar{\rho}_k$ can be written as

$$\bar{\rho}_k = \frac{\gamma \lambda_{\mathrm{u}} L A_k}{\lambda_k W \log_2\left(1+\tau\right)} \cdot \frac{1}{A_k \sum\limits_{j=1}^{K} \tilde{\lambda}_j \tilde{P}_j^{-\frac{2}{\alpha}} \bar{\rho}_j Z + 1}, \tag{2.26}$$

where $k \in \{1, ..., K\}$. It can be seen from (2.26) that the average traffic intensity of one tier is closely related to that of other tiers, which forms the set of fixed-point equations. Although no closed-form expression of $\bar{\rho}_k$ can be obtained directly, the existence and uniqueness of the solution of (2.26) will be proved and an iterative method to solve it will be proposed in Chapter 5 by a 2-Tier HetNet.

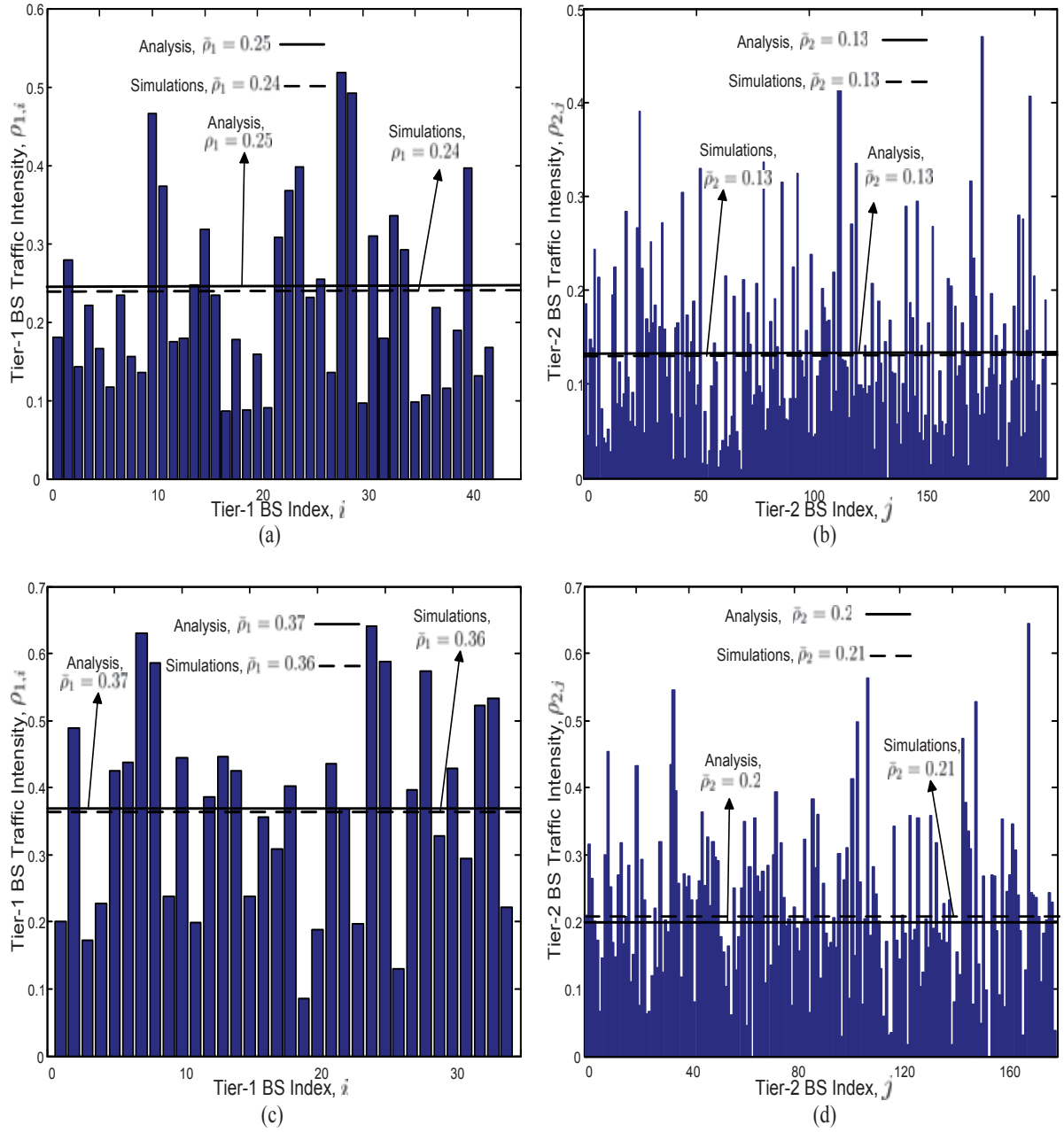


Figure 2.4: Traffic intensity of each BS in one simulation run under the assumption of spectrum partitioning and universal frequency reuse. Note that for spectrum partitioning, the bandwidth is divided as $W_1 = W_2 = 6$MHz and the mean arrival bit rate per area is $\gamma\lambda_{\mathrm{u}}L = 60\mathrm{Mbps/km}^2$; for universal frequency reuse, the mean arrival bit rate per area is $\gamma\lambda_{\mathrm{u}}L = 200\mathrm{Mbps/km}^2$. (a) Tier-1 BS traffic intensity $\rho_{1,i}$ with spectrum partitioning. (b) Tier-2 BS traffic intensity $\rho_{2,j}$ with spectrum partitioning. (c) Tier-1 BS traffic intensity $\rho_{1,i}$ with universal frequency reuse. (b) Tier-2 BS traffic intensity $\rho_{2,j}$ with universal frequency reuse.

Table 2.2: Simulation Parameters

Parameter	Value
Tier-1 BS Transmission Power P_1	20 W
Tier-2 BS Transmission Power P_2	6 W
Path Loss Coefficient α	4
SIR Threshold τ	1
Mean Packet Length L	0.001 Mb
Total Bandwidth W	12 MHz

2.3.3 Simulation Results

In this section, we will justify the proceeding analysis in Section 2.3.1 and Section 2.3.2 by conducting a spatial-temporal simulation of a 2-Tier HetNet. In the simulation, we first deploy BSs of each tier by independent Poisson Point Processes (PPPs) in a square area of $4*10^6\text{m}^2$. The deployment densities of the two tiers are $\lambda_1 = 1*10^{-5}\text{m}^{-2}$ and $\lambda_2 = 5*10^{-5}\text{m}^{-2}$, respectively. Users are deployed according to another PPP with the intensity $\lambda_\text{u} = 10^{-3}\text{m}^{-2}$. Each user then associates to their BSs by the largest average reference signal receiving power. The other system parameters are summarized in Table 2.2.

Fig. 2.4 illustrate the simulation results of the traffic intensity $\rho_{k,i}$ of each Tier-k BS in one simulation run. Note that one simulation run corresponds to one realization of the Poisson point process, and last for 10^4s. It can be observed from Fig. 2.4 that although the traffic intensity $\rho_{k,i}$ of each Tier-k BS varies, the simulation result of the average traffic intensity $\bar{\rho}_k$ over a large region converges to a certain value in different realizations for both spectrum partitioning and universal frequency reuse. This is quite similar to the ergodicity of the Poisson point process, i.e., the spatial averages obtained by averaging over a realization of the PPP over a large region equal the ensemble averages obtained

by averaging over the point process. By extensive simulation runs of different network topology, it turns out that the average traffic intensity $\bar{\rho}_k$ is quite close to the analytical results (2.20) and (2.26), which indicates that $\bar{\rho}_k$ can be well predicted by the independent thinning approach adopted in this thesis.

2.4 Conclusions

This chapter analyzes the queuing performance of the BSs in each tier in heterogeneous networks. In particular, the queuing model and the formulation of the coupled queue problem with fixed locations of BSs and users are first identified. By using stochastic geometry as a tool, the set of BSs of each tier and the set of users are then modeled by independent Poisson point processes, based upon which the average traffic intensity of each tier is characterized. Independent thinning approach is introduced and adopted to obtain the expression of the average traffic intensity of each tier. It is further shown that with the strategy of spectrum partitioning, an closed-form solution of the average traffic intensity is derived; With universal frequency reuse, the average traffic intensity of each tier is determined a set of fixed-point equations. At last, the queuing analysis is justified by a spatial-temporal simulation of a 2-Tier HetNet.

Chapter 3

Queue-Aware Delay-Optimal Biased Association Optimization in HetNets

In this chapter, we will study how to optimally tune the biasing factor of the BSs of each tier to improve the delay performance of a HetNet. Based on the queuing analysis in Chapter 2, the average traffic intensity with respect to the biasing factor of each tier is explicitly derived, and is shown to be an increasing function of the biasing factor. In order to find the delay limit that the network can achieve, an optimization problem is formulated to minimize a lower bound of the network mean queuing delay. By showing that the optimization problem is convex, the optimal biasing factor of each tier can be obtained numerically. When the mean packet arrival rate of each user is small, a closed-form solution is derived. The simulation results demonstrate that the network queuing performance can be significantly improved by properly tuning the biasing factor. It is further shown that the network mean queuing delay might be improved at the cost of a deterioration of the network SIR coverage, which indicates a performance tradeoff between real-time and non-real-time traffic in HetNets.

3.1 Introduction

Among all the techniques used in HetNets, load balancing plays a key role to determine the network performance. For example, to purposely push users to micro BSs, [46, 59, 65–68, 113, 114] proposed a biased association scheme where each user assigned a biased value to the measured received power from BSs of each tier, and associated with the BS with the largest mean biased received power. A detailed review can be found in Section 1.2.1.

However, since the aforementioned studies [46, 59, 65–68, 113, 114] assumed that the BSs always have packets to transmit, they presented a worst case for the performance metrics such as the network SINR and rate coverage, and did not characterize the delay-related performance. One BS, nevertheless, can vary between busy and idle states over a small time scale due to the dynamic packet arrivals of its associated users, in which case the packet delay could be taken into account. In practice, with the proliferation of real-time multimedia applications, the packet delay is becoming an important QoS metric. For example, an end-to-end latency over 200 ms for real-time video media stream is generally considered to be unacceptable [73]. Hence, this motivates us to derive the delay performance metrics analytically with the consideration of queuing, such that some system design insights can be gained to balance the load pressure across tiers.

In particular, we consider a K-Tier HetNet in this chapter where users and BSs of all tiers are randomly distributed, i.e., follow a PPP distribution. Similar to previous studies [59, 65–68, 114], it is assumed that each user adopts a biased association scheme to choose one BS with the maximum biased received power. The packet requests from the users is assumed to form a queue in their associated BSs. In order to find the delay limit that the network can achieve, an optimization problem is formulated to minimize a lower bound of the network mean queuing delay by optimizing over the biasing factor of each tier based on the derived average traffic intensity in Chapter 2. It is shown that the optimization problem is convex, and the optimal biasing factor can be numerically

obtained. When the mean packet arrival rate of each user is small, an explicit expression of the optimal biasing factor of each tier is obtained. With equal bandwidth allocation across tiers, it is further shown that each user should associate with its nearest BS. A case study of a 2-Tier HetNet shows that the optimal biasing factor is sensitive to the bandwidth allocation of each tier. To characterize the network capacity to support non-real-time services, the network SIR coverage is further derived. The contributions of this chapter are summarized as follows.

- By assuming queuing in each BS, an explicit expression of the average traffic intensity of each tier is derived, which is shown to be an increasing function of the biasing factor of each tier.
- An optimization problem of a lower bound of the network mean queuing delay is formulated, and is shown to be convex with respect to the biasing factor of each tier. When the mean packet arrival rate of each user is small, an explicit solution is obtained.
- Simulation results of a 2-tier case demonstrate that the network mean queuing delay can be significantly reduced by properly tuning the biasing factor of each tier. In the meanwhile, a tradeoff is revealed between the network mean queuing delay and the network SIR coverage, which indicates that the service provider should strike a balance between the performance of real-time and non-real-time services.

The rest of this chapter is organized as follows. The system model is presented in Section 3.2. An optimization problem to minimize a lower bound of the network mean queuing delay is formulated and studied in Section 3.3. A case study of a 2-tier HetNet is presented in Section 3.4. Conclusions are given in Section 3.5.

3.2 System Model

Consider a K-tier heterogeneous network where BSs in the kth tier form an independent PPP Φ_k with an intensity of λ_k, $k \in \{1, ..., K\}$. Users, on the other hand, form another independent homogeneous PPP Φ_{u} with an intensity of λ_{u} over the whole network. Frequency partitioning across tiers is assumed in this chapter. In particular, BSs of the same tier share the spectrum with a bandwidth of W_k, $k \in \{1, ..., K\}$, and BSs of different tiers occupy non-overlapping frequency bands. Therefore, for each user in the downlink, the associated BS acts as a desired signal transmitter, and other BSs of the same tier are interfering sources. Consider a typical user located at the origin. Denote the distance between this typical user and a Tier-k BS as x_k, and the transmission power of a Tier-k BS as P_k. The received power P_{R} for a typical user from this BS can then be written as

$$P_{\mathrm{R}} = P_k g_k x_k^{-\alpha}, \tag{3.1}$$

where g_k is a small-scale fading coefficient, which is assumed to follow an i.i.d exponential distribution of unit mean, i.e., $g_k \sim \exp\{1\}$, and α is the path-loss coefficient, which is assumed to be the same for all BSs in the network. Note that shadowing, i.e., log-normal fading, can be modeled by the randomness of the BSs' and users' locations [111].

In this chapter, we consider a biased association scheme where users associate with one BS according to the maximum mean biased received power [59, 65–68, 114]. In particular, for a typical user located at the origin, it measures the mean received power from each tier's BSs, and chooses a Tier-k BS if

$$P_k B_k x_{k,\min}^{-\alpha} \geq P_j B_j x_{j,\min}^{-\alpha} \quad \forall j \in \{1, ..., K\}, \tag{3.2}$$

where B_j denotes the biasing factor of Tier j and $x_{j,\min}$ is the distance between the user and the nearest Tier-j BS.

For each user in the network, assume that its packet requests follow an independent

Poisson process with a mean arrival rate γ, and the packet length is exponentially distributed with mean L. The incoming packets for all users form a queue in the associated BS, and the BS will transmit these packets in a first-in-first-serve (FIFS) fashion. Note that a more complicated scheme can be that the packet is stored in the buffer when the SIR is low and wait for next transmission opportunity when the SIR becomes higher than the threshold. However, as this thesis mainly focuses on the performance of some delay-sensitive applications such as online gaming, this more complicated scheme would result in high queuing delay even for the users that have good channel conditions. Therefore, it is beyond the scope of this thesis. But it still deserves much attention in the future study. To avoid users in poor channel conditions occupying the BS, we consider a fixed rate modulation and coding format. In particular, a BS will serve a user only when its instantaneous SIR exceeds a threshold τ, and will drop its packet request otherwise.

3.3 Queuing Delay Optimization

In this section, we will characterize the minimization problem of the network mean queuing delay by optimizing the association probability (biasing factor) of each tier. As the mean queuing delay of a BS increases with a higher busy probability, we will first study how the average traffic intensity of one tier varies with the association probability of this tier.

3.3.1 Relation Between Average Traffic Intensity and Association Probability

Recall in Section 2.3.1 that the average traffic intensity $\bar{\rho}_k$ of Tier-k BSs can be derived as

$$\bar{\rho}_k = \frac{-\lambda_k W_k \log_2(1+\tau) + \left[(\lambda_k W_k \log_2(1+\tau))^2 + 4\gamma\lambda_u\lambda_k A_k^2 W_k \log_2(1+\tau) LZ\right]^{\frac{1}{2}}}{2A_k\lambda_k Z W_k \log_2(1+\tau)}, \quad (3.3)$$

where Z denotes $Z(\tau,\alpha)$ for simplicity. Note that as a biased association scheme is adopted, the association probability A_k in (3.3) should be modified as

$$A_k=\frac{\lambda_k(P_kB_k)^{2/\alpha}}{\sum\limits_{j=1}^{K}\lambda_j(P_jB_j)^{2/\alpha}}=\frac{1}{\sum\limits_{j=1}^{K}\widetilde{\lambda}_j\left(\widetilde{B}_j\widetilde{P}_j\right)^{2/\alpha}} \tag{3.4}$$

according to [59], where $\widetilde{\lambda}_j=\lambda_j/\lambda_k$, $\widetilde{P}_j=P_j/P_k$, and $\widetilde{B}_j=B_j/B_k$ denote the normalized intensity, the normalized transmission power, and the normalized biasing factor of Tier j, respectively, conditioned on Tier k being a serving tier. As it can be easily observed from (3.4) that the association probability $\{A_k\}_{\forall k}$ is uniquely determined by the biasing factor $\{B_k\}_{\forall k}$, we will optimize A_k instead of optimizing B_k of each tier in the rest of this chapter.

It is indicated in (3.3) that $\bar{\rho}_k$ is critically determined by the mean packet arrival rate of each user γ and the association probability A_k. It is clear that $\bar{\rho}_k$ increases as γ increases. On the other hand, the following lemma presents the monotonicity of the average traffic intensity $\bar{\rho}_k$ of Tier-k BSs with respect to the association probability A_k.

Lemma 3.1. *The average traffic intensity $\bar{\rho}_k$ of Tier-k BSs is an increasing function of its association probability, A_k.*

Proof. According to (3.3), the first-order derivative of the average traffic intensity $\bar{\rho}_k$ with respect to A_k is given by

$$\frac{d\bar{\rho}_k}{dA_k}=\frac{4\gamma L\lambda_{\mathrm{u}}\lambda_k^2W_k\log_2(1+\tau)A_k^2Z^2\Delta^{-\frac{1}{2}}-\lambda_kZ\left(-\lambda_kW_k\log_2(1+\tau)+\Delta^{\frac{1}{2}}\right)}{2W_k\log_2(1+\tau)\left(A_k\lambda_kZ\right)^2}, \tag{3.5}$$

where $\Delta=\lambda_k^2W_k^2\log_2^2(1+\tau)+4\gamma\lambda_{\mathrm{u}}\lambda_kW_k\log_2(1+\tau)A_k^2LZ$. The numerator on the right

hand side of (3.5) can be further written as

$$\begin{aligned}&4\gamma L\lambda_{\mathrm{u}}\lambda_k^2 W_k^2 \log_2^2(1+\tau)A_k^2 Z^2 \Delta^{-\frac{1}{2}} - \lambda_k W_k \log_2(1+\tau) Z\left(-\lambda_k W_k \log_2(1+\tau) + \Delta^{\frac{1}{2}}\right)\\ =&\frac{\lambda_k^2 R_k^2 Z\left[\left(\lambda_k^2 W_k^2 \log_2^2(1+\tau) + 4\gamma\lambda_{\mathrm{u}}\lambda_k W_k \log_2(1+\tau)A_k^2 LZ\right)^{\frac{1}{2}} - \lambda_k W_k \log_2(1+\tau)\right]}{\Delta^{\frac{1}{2}}} > 0.\end{aligned} \tag{3.6}$$

By combining (3.5) and (3.6), we have $\frac{d\bar{\rho}_k}{dA_k} > 0$, which indicates that $\bar{\rho}_k$ monotonically increases as A_k increases. □

Intuitively, as the probability of a user being associated with a Tier-k BS increases, more users from other tiers will be offloaded to BSs of Tier k, which leads to an increment of the traffic intensity.

When the mean packet arrival rate of each user γ is small, the average traffic intensity $\bar{\rho}_k$ of Tier-k BSs can be approximately written as

$$\begin{aligned}\bar{\rho}_k &= \frac{-1 + \left[1 + 4\gamma\lambda_{\mathrm{u}} A_k^2 (\lambda_k W_k \log_2(1+\tau))^{-1} LZ\right]^{\frac{1}{2}}}{2A_k Z}\\ &\overset{(a)}{\approx} \frac{-1 + 1 + 2\gamma\lambda_{\mathrm{u}} A_k^2 (\lambda_k W_k \log_2(1+\tau))^{-1} LZ}{2A_k Z}\\ &= \frac{\gamma\lambda_{\mathrm{u}} L A_k}{\lambda_k W_k \log_2(1+\tau)},\end{aligned} \tag{3.7}$$

where (a) follows from the fact that

$$\left[1 + \frac{4\gamma\lambda_{\mathrm{u}} A_k^2 LZ}{\lambda_k W_k \log_2(1+\tau)}\right]^{\frac{1}{2}} \approx 1 + \frac{2\gamma\lambda_{\mathrm{u}} A_k^2 LZ}{\lambda_k W_k \log_2(1+\tau)}. \tag{3.8}$$

Note that since (3.8) becomes more accurate as $\frac{4\gamma\lambda_{\mathrm{u}} A_k^2 LZ}{\lambda_k W_k \log_2(1+\tau)}$ approaches zero, using (3.7) to represent $\bar{\rho}_k$ achieves better approximation with a lower value of the mean packet arrival rate γ, which indicates a network with a lower traffic load pressure.

3.3.2 Queuing Delay Optimization

As each BS can be modeled as a M/D/1 queuing system, the mean queuing delay D_k of Tier k BSs can be obtained as

$$D_k = E\left[\frac{L}{W_k \log_2(1+\tau)(1-\rho_{k,i})}\right]. \tag{3.9}$$

Here we would like to make it clear that as we study the average queuing conditions of the BSs of each tier and each BS is modeled as M/D/1 system, therefore, each Tier-k BS has a traffic intensity and a mean queuing delay which is averaged over time, respectively. From this approach, the average traffic intensity $\bar{\rho}_k$ and the mean queuing delay of a tier D_k is then averaged over all the BSs, respectively, in this tier.

Since (3.9) is difficult to characterize, we resort to its lower bound using the convexity of $1/(1-\rho_{k,i})$, i.e., we have

$$D_k \geq \bar{D}_k = \frac{L}{W_k \log_2(1+\tau)(1-E[\rho_{k,i}])} = \frac{L}{W_k \log_2(1+\tau)(1-\bar{\rho}_k)}. \tag{3.10}$$

By combining (3.3) and (3.10), the lower bound of the mean queuing delay of the whole network $\bar{D}$ can then be written as

$$\bar{D} = \sum_{k=1}^{K} \frac{\lambda_k}{\sum_{j=1}^{K} \lambda_j} \cdot \bar{D}_k = \sum_{k=1}^{K} \frac{2A_k\lambda_k^2 LZ}{\sum_{j=1}^{K} \lambda_j \left(2A_k\lambda_k Z R_k + \lambda_k R_k - \left[(\lambda_k R_k)^2 + 4\gamma\lambda_u\lambda_k A_k^2 R_k LZ\right]^{\frac{1}{2}}\right)}, \tag{3.11}$$

where $R_k = W_k \log_2(1+\tau)$.

It can be observed from (3.11) that the lower bound of the mean queuing delay $\bar{D}$ is critically determined by the association probability A_k. To minimize $\bar{D}$, we have the following optimization problem

$$\underset{\{A_k\}_{\forall k \in \{1,\ldots,K\}}}{\text{minimize}} \quad \bar{D}, \tag{3.12a}$$

$$\text{s.t.} \sum_{k=1}^{K} A_k = 1, \tag{3.12b}$$

$$\bar{\rho}_k < 1,\ k \in \{1, \ldots, K\}. \tag{3.12c}$$

Note that as we optimize over the association probabilities $\{A_k\}_{\forall k}$ in (3.12) to obtain the optimal solution $\{A_k^*\}_{\forall k}$, the optimal normalized biasing factor of Tier k conditioned on Tier i, $\{\widetilde{B}_k^*\}_{\forall k}$, can then be readily obtained as

$$\widetilde{B}_k^* = \frac{P_i(\lambda_i A_k^*)^{\frac{\alpha}{2}}}{P_k(\lambda_k A_i^*)^{\frac{\alpha}{2}}},\ k \in \{1, ..., K\}, \tag{3.13}$$

according to (3.4). On the other hand, the constraint (3.12b) comes from the fact that each user should associate with one BS, and the constraint (3.12c) ensures that the lower bound of the network's mean queuing delay is bounded, which leads to the following lemma.

Lemma 3.2. *For the lower bound $\bar{D}_k$, when the mean packet arrival rate $\gamma > \frac{(Z+1)\lambda_k R_k}{\lambda_u L}$, it is bounded if the association probability*

$$0 < A_k < \frac{\lambda_k R_k}{\gamma \lambda_{\mathrm{u}} L - \lambda_k R_k Z}; \tag{3.14}$$

otherwise, it will become unbounded. When $\gamma < \frac{(Z+1)\lambda_k R_k}{\lambda_u L}$, it is always bounded.

Proof. It has been shown in Lemma 3.1 that the average traffic intensity $\bar{\rho}_k$ monotonically increases as the association probability A_k increases. With $A_k < 1$, we then have

$$\begin{aligned} \bar{\rho}_k &= \frac{-\lambda_k R_k + \left[(\lambda_k R_k)^2 + 4\gamma\lambda_{\mathrm{u}}\lambda_k R_k A_k^2 LZ\right]^{\frac{1}{2}}}{2A_k\lambda_k R_k Z} \\ &< \frac{-\lambda_k R_k + \left[(\lambda_k R_k)^2 + 4\gamma\lambda_{\mathrm{u}}\lambda_k R_k LZ\right]^{\frac{1}{2}}}{2\lambda_k R_k Z}. \end{aligned} \tag{3.15}$$

In the following, we divide the discussion into two parts:

1) If $\frac{-\lambda_k R_k + \left[(\lambda_k R_k)^2 + 4\gamma\lambda_{\mathrm{u}}\lambda_k R_k LZ\right]^{\frac{1}{2}}}{2\lambda_k R_k Z} < 1$, i.e., $\gamma < \frac{(Z+1)\lambda_k R_k}{\lambda_{\mathrm{u}} L}$, we have

$$\bar{\rho}_k < 1 \tag{3.16}$$

according to (3.15). In this case, $\bar{D}_k$ will always be bounded if $\gamma < \frac{(Z+1)\lambda_k R_k}{\lambda_{\mathrm{u}} L}$.

2) If $\gamma > \frac{(Z+1)\lambda_k R_k}{\lambda_u L}$, $\bar{D}_k$ will be bounded if and only if

$$\frac{-\lambda_k R_k + \left[(\lambda_k R_k)^2 + 4\gamma\lambda_{\mathrm{u}}\lambda_k A_k^2 R_k L Z\right]^{\frac{1}{2}}}{2A_k\lambda_k R_k Z} < 1. \tag{3.17}$$

Accordingly, we have

$$A_k < \frac{\lambda_k R_k}{\gamma\lambda_{\mathrm{u}} L - \lambda_k R_k Z}. \tag{3.18}$$

□

According to Lemma 3.2, constraint (3.12c) can be further written as

$$\begin{cases} 0 < A_k < \frac{\lambda_k R_k}{\gamma\lambda_{\mathrm{u}} L - \lambda_k R_k Z}, & \gamma > \frac{(Z+1)\lambda_k R_k}{\lambda_u L} \\ 0 < A_k < 1, & \gamma < \frac{(Z+1)\lambda_k R_k}{\lambda_{\mathrm{u}} L} \end{cases}, \tag{3.19}$$

where $k \in \{1, \ldots, K\}$. First note that (3.19) does not have a feasible region if and only if

$$\gamma > \max_{\forall k} \left\{ \frac{(Z+1)\,\lambda_k R_k}{\lambda_{\mathrm{u}} L} \right\} \tag{3.20a}$$

and

$$\sum_{k=1}^{K} \frac{\lambda_k R_k}{\gamma\lambda_{\mathrm{u}} L - \lambda_k R_k Z} < 1, \tag{3.20b}$$

according to (3.19). Intuitively, when the mean packet arrival rate of each user γ is too large, (3.19) can be written as $0 < A_k < \frac{\lambda_k R_k}{\gamma\lambda_{\mathrm{u}} L - \lambda_k R_k Z}$ for each Tier k, $k \in \{1, \ldots, K\}$, which leads to $\sum_{k=1}^{K} A_k < 1$ according to (3.20b). In this case, the lower bound of the network mean queuing delay will always be unbounded. If (3.20) does not hold, the feasible region of the optimization problem (3.12) can be further written as

$$\begin{aligned} \mathbf{A} = \Bigg\{ & (A_1, ..., A_{K-1}) \,, \Big| 0 < A_j < \min\left\{1, \frac{\lambda_j R_j}{\gamma\lambda_{\mathrm{u}} L - \lambda_j R_j Z}\right\},\ j \in \{1..., K-1\}; \\ & \max\left\{0, 1 - \frac{\lambda_K R_K}{\gamma\lambda_{\mathrm{u}} L - \lambda_K R_K Z}\right\} < \sum_{j=1}^{K-1} A_j \ < 1 \Bigg\}, \end{aligned} \tag{3.21}$$

where A_K is eliminated according to the constraint (3.12b) without loss of generality. The following lemma proves that the objective function (3.12a) is convex within the feasible region **A**

Lemma 3.3. *The objective function* (3.12a) *is convex with respect to the association probability* A_k *within the constraints* (3.12b) *and* (3.12c).

Proof. According to (3.10), the second-order derivative of $\bar{D}_k$ with respect to A_k can be written as

$$\frac{d^2\bar{D}_k}{dA_k{}^2} = \frac{2L}{R_k\left(1-\bar{\rho}_k\right)^3}\cdot\left(\frac{d\bar{\rho}_k}{dA_k}\right)^2 + \frac{L}{R_k\left(1-\bar{\rho}_k\right)^2}\cdot\frac{d^2\bar{\rho}_k}{dA_k{}^2}. \tag{3.22}$$

Substituting (3.5) into (3.22) yields

$$\begin{aligned}
&\frac{d^2\bar{D}_k}{dA_k{}^2} > \frac{L}{R_k\left(1-\bar{\rho}_k\right)^2}\cdot\left[2\left(\frac{d\bar{\rho}_k}{dA_k}\right)^2 + \frac{d^2\bar{\rho}_k}{dA_k{}^2}\right] = \frac{L}{R_k\left(1-\bar{\rho}_k\right)^2 A_k^4 Z^2\Delta}\\
&\cdot\left(4\gamma\lambda_u L\lambda_k^2 R_k^2 Z^2 A_k^3 + 2\Delta + 2\lambda_k R_k A_k Z\Delta^{\frac{1}{2}} + \lambda_k^2 R_k^2 - 2A_k Z\Delta - \lambda_k R_k\Delta^{\frac{1}{2}}\right)\\
&> \frac{L}{R_k\left(1-\rho_k\right)^2 A_k^4 Z^2\Delta}\cdot\left[4\gamma\lambda_u L\lambda_k^2 R_k^2 Z^2 A_k^3 + \lambda_k R_k\left(2A_k Z\Delta^{\frac{1}{2}} + \lambda_k R_k - \Delta^{\frac{1}{2}}\right)\right],
\end{aligned} \tag{3.23}$$

where $\Delta = \lambda_k^2 R_k^2 + 4\gamma\lambda_{\mathrm{u}}\lambda_k R_k A_k^2 LZ$. Since $\Delta^{\frac{1}{2}} > \lambda_k R_k$, we further have

$$\begin{aligned}
\frac{d^2\bar{D}_k}{dA_k{}^2} &> \frac{L}{R_k\left(1-\rho_k\right)^2 A_k^4 Z^2\Delta}\cdot\left[4\gamma\lambda_u L\lambda_k^2 R_k^2 Z^2 A_k^3 + \lambda_k R_k\left(2A_k Z\Delta^{\frac{1}{2}} + \lambda_k R_k - \Delta^{\frac{1}{2}}\right)\right]\\
&> \frac{L}{R_k\left(1-\rho_k\right)^2 A_k^4 Z^2\Delta}\cdot\left[4\gamma\lambda_u L\lambda_k^2 R_k^2 Z^2 A_k^3 + \lambda_k R_k\left(2\lambda_k R_k A_k Z + \lambda_k R_k - \Delta^{\frac{1}{2}}\right)\right]\\
&\overset{(a)}{>} \frac{4\gamma\lambda_u L^2\lambda_k^2 R_k}{\left(1-\rho_k\right)^2 A_k\Delta} > 0,
\end{aligned} \tag{3.24}$$

where (a) follows from the fact that $\bar{\rho}_k < 1$, As the constraints (3.12b) and (3.12c) are linear, it can be concluded from (3.24) that the optimization problem is convex with respect to A_k. □

Nevertheless, there may not exist a solution in **A** by setting the partial derivative of $\bar{D}$

with respect to the association probability A_k to zero, i.e.

$$\frac{\partial \bar{D}}{\partial A_k} = -2\lambda_k Z \frac{R_k A_k^{-2} - R_k^2 A_k^{-3}\left(R_k^2 A_k^{-2} + 4\gamma\lambda_{\mathrm{u}}\lambda_k^{-1} R_k L Z\right)^{-\frac{1}{2}}}{\left[2ZR_k + R_k A_k^{-1} - \left(R_k^2 A_k^{-2} + 4\gamma\lambda_{\mathrm{u}}\lambda_k^{-1} R_k L Z\right)^{\frac{1}{2}}\right]^2} + 2\lambda_K Z$$
$$\times \frac{R_K\left(1 - \sum_{j=1}^{K-1} A_j\right)^{-2} - R_K^2\left(1 - \sum_{j=1}^{K-1} A_j\right)^{-3}\left[R_K^2\left(1 - \sum_{j=1}^{K-1} A_j\right)^{-2} + 4\gamma\lambda_{\mathrm{u}}\lambda_K^{-1} R_K L Z\right]^{-\frac{1}{2}}}{\left\{2ZR_K + R_K\left(1 - \sum_{j=1}^{K-1} A_j\right)^{-1} - \left[R_K^2\left(1 - \sum_{j=1}^{K-1} A_j\right)^{-2} + 4\gamma\lambda_{\mathrm{u}}\lambda_K^{-1} R_K L Z\right]^{\frac{1}{2}}\right\}^2}$$
$$= 0, \quad k \in \{1, \ldots, K-1\}. \tag{3.25}$$

The following lemma rules out this possibility and guarantees that the optimal association probabilities $\{A_k^*\}_{\forall k}$ can always be obtained by finding the solution of (3.25) within $\mathbf{A}$.

Lemma 3.4. (3.25) *has a unique solution within the feasible region* $\boldsymbol{A}$*, which is the optimal association probabilities* $\{A_k^*\}_{\forall k}$*.*

Proof. We divide the proof into two parts.

1) If $\gamma > \max\limits_{\forall k}\left\{\frac{(Z+1)\lambda_k R_k}{\lambda_{\mathrm{u}} L}\right\}$, then the mean queuing delay $\bar{D}_k$ of all tiers go to infinity as A_k approaches to 1. Therefore, according to (3.11), the lower bound of the network mean queuing delay, $\bar{D}$, goes to infinity at the boundary of $\mathbf{A}$. As $\bar{D}$ is convex within the region $\mathbf{A}$, (3.25) always has a unique solution of the optimal association probabilities $\{A_k^*\}_{\forall k}$.

2) If $\gamma < \max\limits_{\forall k}\left\{\frac{(Z+1)\lambda_k R_k}{\lambda_{\mathrm{u}} L}\right\}$, then there exists at least one tier such that the lower bound of its mean queuing delay is always bounded. Without loss of generality, denote this tier as Tier K. For Tier K, we have $\frac{\lambda_K R_K}{\gamma\lambda_{\mathrm{u}} L - \lambda_K R_K Z} > 1$, and the feasible region $\mathbf{A}$ is then written as

$$\mathbf{A} = \Bigg\{ (A_1, ..., A_{K-1}) \,\Bigg|\, 0 < A_k < \min\left\{1, \frac{\lambda_k R_k}{\gamma\lambda_{\mathrm{u}} L - \lambda_k R_k Z}\right\},\ k \in \{1..., K-1\};$$
$$0 < \sum_{k=1}^{K-1} A_k < 1 \Bigg\}. \tag{3.26}$$

For each $k \in \{1, \ldots, K-1\}$, we have

$$\lim_{A_k \to 0} \frac{\partial \bar{D}}{\partial A_k} = 2\lambda_K Z \cdot \frac{R_K A_K^{-2} \left[1 - \left(1+4\gamma\lambda_{\mathrm{u}}\lambda_K^{-1} A_K^2 R_K^{-1} LZ\right)^{-\frac{1}{2}}\right]}{\left[2ZR_K + R_K A_K^{-1} - \left(R_K^2 A_K^{-2} + 4\gamma\lambda_{\mathrm{u}}\lambda_k^{-1} R_k LZ\right)^{\frac{1}{2}}\right]^2} < 0 \tag{3.27}$$

according to (3.25).

Following a similar approach, if $\frac{\lambda_k R_k}{\gamma\lambda_{\mathrm{u}} L - \lambda_k R_k Z} > 1$, we have

$$\lim_{A_k \to 1} \frac{\partial \bar{D}}{\partial A_k} > 0. \tag{3.28}$$

Otherwise, if $\frac{\lambda_k R_k}{\gamma\lambda_{\mathrm{u}} L - \lambda_k R_k Z} < 1$, the lower bound $\bar{D}_k$ goes to infinity as A_k approaches $\frac{\lambda_k R_k}{\gamma\lambda_{\mathrm{u}} L - \lambda_k R_k Z}$, and thus we have

$$\lim_{A_k \to \frac{\lambda_k R_k}{\gamma\lambda_u L - \lambda_k R_k Z}} \frac{\partial \bar{D}}{\partial A_k} > 0. \tag{3.29}$$

By combining (3.27), (3.28) and (3.29), it can be concluded that (3.25) always has only one solution within the region $0 < A_k < \min\{1, \frac{\lambda_k R_k}{\gamma\lambda_{\mathrm{u}} L - \lambda_k R_k Z}\}$, $k \in \{1, \ldots, K-1\}$.

Furthermore, if $\sum_{k=1}^{K-1} A_k > 1$, i.e., $A_K < 0$, we always have $\frac{\partial \bar{D}}{\partial A_k} > 0$, $k \in \{1, \ldots, K-1\}$ by substituting $A_K < 0$ into (3.25). This indicates that the solution is not in the region where $\sum_{k=1}^{K-1} A_k > 1$. Therefore, (3.25) has a unique solution in region **A** when $\gamma < \max_{\forall k} \left\{ \frac{(Z+1)\lambda_k R_k}{\lambda_{\mathrm{u}} L} \right\}$. □

So far we have demonstrated how to find the optimal association probability of each tier A_k^* by solving (3.25) numerically. Recall that it is indicated in Lemma 3.2 that when the mean packet arrival rate of each user $\gamma < \min_{\forall k} \left\{ \frac{(Z+1)\lambda_k R_k}{\lambda_{\mathrm{u}} L} \right\}$, we have the average traffic intensity $\bar{\rho}_k < 1$ for all tiers, and the lower bound of the mean queuing delay $\bar{D}_k$ is always bounded for each tier. In this case, the average traffic intensity $\bar{\rho}_k$ is simply written as (3.7), and an explicit optimal association probability A_k^* for each tier can be obtained, which is shown in the following lemma.

Lemma 3.5. *When the mean packet arrival rate of each user* $\gamma < \min\limits_{\forall k}\left\{\frac{(Z+1)\lambda_k R_k}{\lambda_u L}\right\}$*, the optimal association probability of Tier* k*,* A_k^**, to minimize the lower bound of the network mean queuing delay* $\bar{D}$ *can be written as*

$$A_k^* = \frac{\lambda_k}{\sum\limits_{j=1}^{K}\lambda_j} + \frac{\lambda_k \log_2(1+\tau)\sum\limits_{j=1}^{K}\lambda_j(W_k - W_j)}{\gamma\lambda_{\mathrm{u}} L \sum\limits_{j=1}^{K}\lambda_j}. \tag{3.30}$$

Proof. By combining (3.7), (3.11), and (3.12b), when the mean packet arrival rate of each user satisfies $\gamma < \min\limits_{\forall k}\left\{\frac{(Z+1)\lambda_k R_k}{\lambda_{\mathrm{u}} L}\right\}$, the lower bound of the network mean queuing delay can be written as

$$\bar{D} = \frac{1}{\sum\limits_{j=1}^{K}\lambda_j}\sum_{k=1}^{K}\frac{\lambda_k^2 L}{\lambda_k R_k - \gamma\lambda_{\mathrm{u}} L A_k} = \frac{1}{\sum\limits_{j=1}^{K}\lambda_j}$$

$$\cdot\left[\sum_{k=1}^{K-1}\frac{\lambda_k^2 L}{\lambda_k R_k - \gamma\lambda_{\mathrm{u}} L A_k} + \frac{\lambda_K^2 L}{\lambda_K R_K - \gamma\lambda_{\mathrm{u}} L(1-\sum\limits_{j=1}^{K-1} A_j)}\right]. \tag{3.31}$$

By setting the partial derivative of $\bar{D}$ with respect to A_k to zero, we have

$$\frac{\partial \bar{D}}{\partial A_k} = \frac{\lambda_k^2}{\sum\limits_{j=1}^{K}\lambda_j}\cdot\frac{\lambda_{\mathrm{u}}\gamma}{\left(\lambda_k\frac{R_k}{L} - \lambda_{\mathrm{u}}\gamma A_k\right)^2} - \frac{\lambda_K^2}{\sum\limits_{j=1}^{K}\lambda_j}\cdot\frac{\lambda_{\mathrm{u}}\gamma}{\left[\lambda_k\frac{R_k}{L} - \lambda_{\mathrm{u}}\gamma\left(1-\sum\limits_{j=1}^{K-1}A_j\right)\right]^2}$$

$$= 0,\ \forall k \in \{1,\ldots,K-1\}. \tag{3.32}$$

By combining (3.12b) and (3.32), (3.30) can be obtained. □

Intuitively, if the bandwidth of Tier k is larger than that of Tier j, i.e., $W_k > W_j$, the service rate of Tier k will be larger, indicating a better queuing performance. Therefore, the Tier-k BSs will undertake more traffic from other tiers by having a larger association

Algorithm 1 Procedure to optimize the association probability when the mean packet arrival rate of each user $\gamma < \min\limits_{\forall k} \left\{ \frac{(Z+1)\lambda_k R_k}{\lambda_{\mathrm{u}} L} \right\}$

1: **Input:** λ_k, W_k for each tier, and other system parameters
λ_{u}, L, γ, τ.

2: **Initialize:** a set of index $C = \{1, \ldots, K\}$ where optimal
association probability of Tier k is not determined.

3: Calculate the solution set $\{A_k^*\}_{\forall k \in C}$ by (3.30).

4: **for** $\forall k \in C$, construct a set $S = \{m \,|A_m^* < 0, \forall m \in C\}$.

5: **if** $S = \emptyset$, **return** $\{A_k^*\}_{\forall k \in C}$.

6: **else**, **for** $\forall m \in S$, let $\lambda_m = 0$ and $A_m^* = 0$, delete m from C.

7: **end if**

8: **go to** Step 3.

probability. With equal bandwidth allocation among all tiers, i.e., $W_i = W_j$, $i, k \in \{1, \ldots, K\}$, the optimal association probability of a Tier-k BS can be further written as

$$A_k^* = \frac{\lambda_k}{\sum\limits_{j=1}^{K} \lambda_j} \tag{3.33}$$

according to (3.30). The corresponding optimal normalized biasing factor $\widetilde{B}_k^*$ of Tier k, conditioned on Tier i, is thus given by

$$\widetilde{B}_k^* = \frac{1}{\widetilde{P}_k}, \tag{3.34}$$

where $\widetilde{P}_k$ is the normalized transmission power of Tier k conditioned on Tier i. It is indicated in (3.34) that in this case, each user chooses the nearest BS. The traffic load is thus evenly distributed among all BSs, which leads to similar queuing performance with the same service rate of each tier's BSs.

Table 3.1: Simulation Parameters

Parameter	Value
User Density λ_{u}	10^{-3} m^{-2}
Tier-1 BS Density λ_1	10^{-5} m^{-2}
Tier-2 BS Density λ_2	$5{*}10^{-5}$ m^{-2}
Tier-1 BS Transmission Power P_1	40 W
Tier-2 BS Transmission Power P_2	3 W
Path Loss Coefficient α	4
Mean Packet Length L	0.001 Mb

Note from (3.30) that if there exists one tier, say Tier m, such that

$$\log_2(1{+}\tau)\sum_{j=1}^{K}\lambda_j(R_m{-}R_j) < -\gamma\lambda_u L, \tag{3.35}$$

and then we have $A_m^* < 0$. To minimize the lower bound of the network mean queuing delay, the association probability of Tier m should be close to zero. Intuitively, if the bandwidth of Tier m is much smaller than that of other tiers, then few users should associate with Tier-m BSs due to the low service rate. In this case, the association probability A_m could then be simply set as $A_m = 0$, i.e., Tier-m BSs are turned off. The procedure to obtain the optimal association probability when $\gamma < \min\limits_{\forall k}\left\{\frac{(Z+1)\lambda_k R_k}{\lambda_{\mathrm{u}} L}\right\}$ is summarized in Algorithm 1.

3.4 Simulation Results

In this section we will validate the analytical results by simulations of a 2-Tier HetNet. The base stations and the users are drawn from PPPs with high intensities, and the background noise is ignored in the simulations. This setting, for example, can correspond to a dense heterogeneous network that consists of macro cellular BSs and micro Wi-Fi

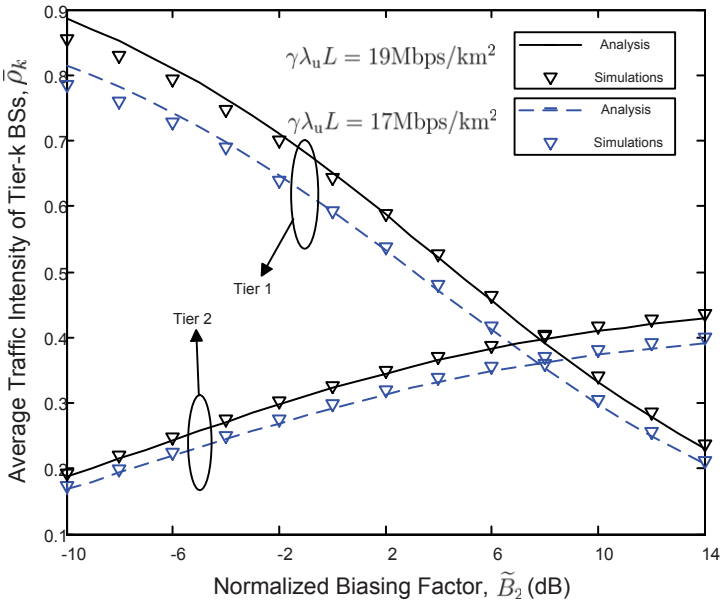


Figure 3.1: Average traffic intensity of each tier $\bar{\rho}_k$ versus the normalized biasing factor $\widetilde{B}_2$ with various values of the mean arrival bit rate per area $\gamma\lambda_u L$. $W_1 = 10$MHz, $W_2 = 6$MHz, and $\tau = 1$.

access points, each of which uses a non-overlapping frequency band. Each point of the simulation results is obtained by averaging all the BSs on a time scale of 10^5s. The system parameters used in simulations are summarized in Table 3.1.

Fig. 3.1 illustrates how the average traffic intensity of each tier, i.e., $\bar{\rho}_1$ and $\bar{\rho}_2$, varies with the normalized biasing factor $\widetilde{B}_2$ with various values of the mean arrival bit rate per area $\gamma\lambda_u L$. It can be observed from Fig. 3.1 that the average traffic intensity of Tier 1, $\bar{\rho}_1$, decreases as the normalized biasing factor $\widetilde{B}_2$ increases, while that of Tier 2, $\bar{\rho}_2$, increases. Intuitively, since the association probability of a Tier-2 BS, A_2, increases as the normalized biasing factor $\widetilde{B}_2$ increases according to (3.4), more users that associate with Tier-1 BSs would be offloaded to Tier-2 BSs, which leads to an increment of $\bar{\rho}_2$ according to Lemma 2. Moreover, due to a larger deployment intensity of the Tier-2 BSs, the users that originally associate with only one Tier-1 BS can be offloaded to several neighboring Tier-2 BSs. Hence, the decline rate of $\bar{\rho}_1$ is larger than the increasing rate of $\bar{\rho}_2$. It can be

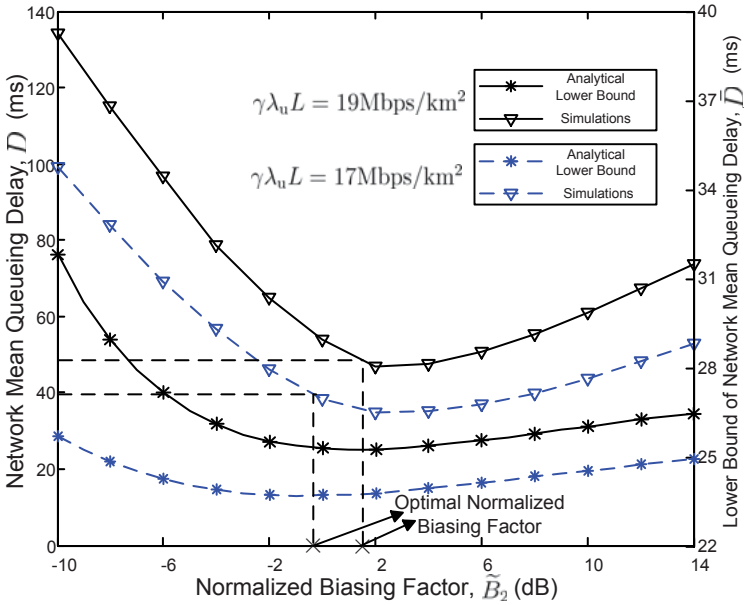


Figure 3.2: Network mean queuing delay D and its lower bound $\bar{D}$ versus the normalized biasing factor $\widetilde{B}_2$ with various values of the mean arrival bit rate per area $\gamma\lambda_u L$. $W_1 = 10$MHz, $W_2 = 6$MHz, and $\tau = 1$.

clearly seen from Fig. 3.1 that the simulation results match with the analysis well with a wide range of the normalized biasing factor, indicating that the independent thinning by replacing each BS's traffic intensity by the average traffic intensity in (2.11) achieves a good approximation.

Fig. 3.2 further demonstrates how the network mean queuing delay D, as well as its lower bound $\bar{D}$, vary with the normalized biasing factor $\widetilde{B}_2$. For the sake of comparison, the y-axis on the left hand side of Fig. 3.2 denotes the network mean queuing delay D while on the right hand side it denotes the lower bound $\bar{D}$. To obtain the network mean queuing delay in simulations, BSs that have an unbounded queuing delay are not taken account of. It can be observed from Fig. 3.2 that the trend of the network mean queuing delay D resembles that of its lower bound $\bar{D}$. Both D and $\bar{D}$ are very sensitive to the normalized biasing factor $\widetilde{B}_2$. If $\widetilde{B}_2$ is not carefully tuned, the delay performance could be greatly degraded. For example, when $\gamma\lambda_u L = 19$Mbps/km^2, the network mean queuing

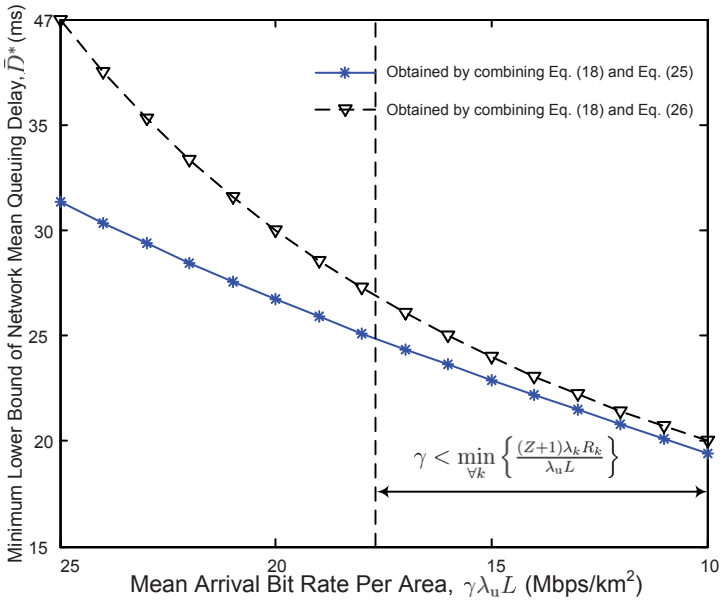


Figure 3.3: Minimum lower bound of the network mean queuing delay $\bar{D}^*$ versus the mean arrival bit rate per area $\gamma\lambda_{\rm u}L$. $W_1 = 10$MHz, $W_2 = 6$MHz, and $\tau = 1$.

delay D is as high as 135ms with the normalized biasing factor $\widetilde{B}_2 = -10$dB, which is not acceptable to many delay-sensitive applications. Moreover, due to a similar trend between the network mean queuing delay D and its lower bound $\bar{D}$, the optimal normalized biasing factor of $\bar{D}$ is close to that of D. Therefore, by properly tuning the normalized biasing factor $\widetilde{B}_2$ according to (3.13) and (3.25), the mean queuing delay performance can be improved significantly. With the mean arrival bit rate per area $\gamma\lambda_{\rm u}L = 19$Mbps/km^2, for instance, the optimal normalized biasing factor is obtained as $\widetilde{B_2^*} = 1.7$dB, and the corresponding network mean queuing delay D can be reduced to be 48ms.

Recall that it is indicated in Lemma 3.5 that when the mean packet arrival rate satisfies $\gamma < \min\limits_{\forall k} \left\{ \frac{(Z+1)\lambda_k R_k}{\lambda_{\rm u} L} \right\}$, the minimum lower bound of the network mean queuing delay $\bar{D}^*$ can be obtained by combining (3.11) and (3.30). Fig. 3.3 further compares the minimum lower bound of the network mean queuing delay $\bar{D}^*$ obtained by combining (3.11) and (3.25) with that by combining (3.11) and (3.30), respectively. It can be observed from Fig. 3.3 that the gap between the two curves diminishes as the mean arrival bit

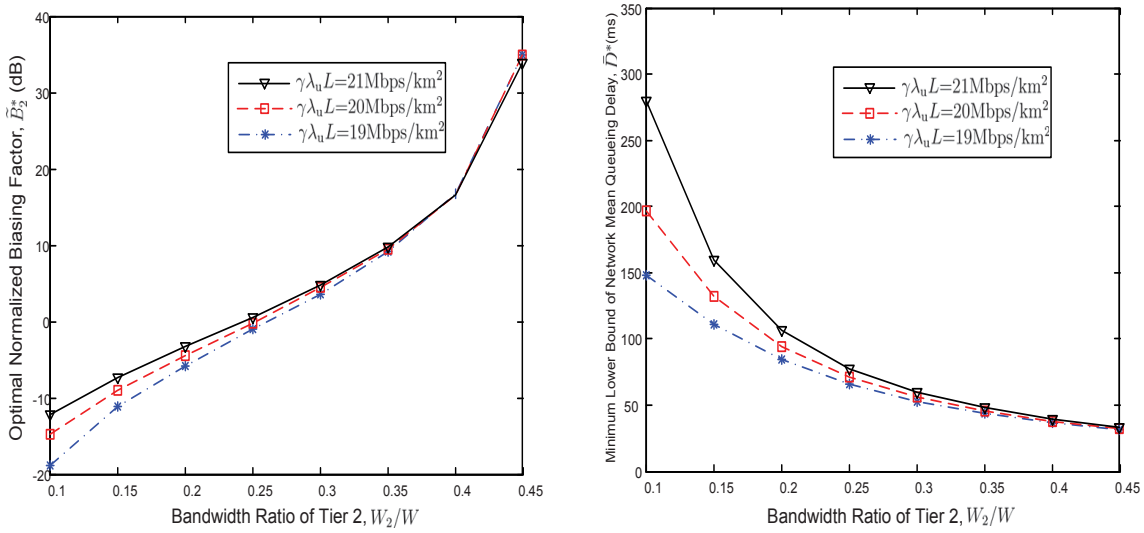
rate per area $\gamma\lambda_{\mathrm{u}}L$ decreases. When the mean arrival bit per area $\gamma\lambda_{\mathrm{u}}L$ is small, i.e., $\gamma < \min\limits_{\forall k}\left\{\frac{(Z+1)\lambda_k R_k}{\lambda_{\mathrm{u}}L}\right\}$, the minimum lower bound of the network mean queuing delay $\bar{D}^*$ obtained by combining (3.11) and (3.30) is quite close to that obtained by combining (3.11) and (3.25). When the mean arrival bit per area $\gamma\lambda_{\mathrm{u}}L$ is large, i.e., $\gamma \geq \min\limits_{\forall k}\left\{\frac{(Z+1)\lambda_k R_k}{\lambda_{\mathrm{u}}L}\right\}$, there is a large gap between the curves in Fig. 3.3. Therefore, the optimal association probabilities $\{A_k^*\}_{\forall k}$ should be instead obtained by numerically solving (3.25). As Lemma 3.4 guarantees, (3.25) has a unique solution within the feasible region $\mathbf{A}$, which is the optimal association probability $\{A_k^*\}_{\forall k}$.

Fig. 3.4 further illustrates how the optimal normalized biasing factor, $\widetilde{B_2^*}$, and the corresponding minimum lower bound of the network mean queuing delay, $\bar{D}^*$, vary with the bandwidth ratio of Tier 2, W_2/W, with various values of the mean packet arrival rate of each user γ. Note that the total bandwidth $W = W_1 + W_2$ is fixed here. It can be observed from Fig. 3.4(a) that for a given γ, the optimal normalized biasing factor $\widetilde{B_2^*}$ increases as W_2/W increases. Intuitively, as the bandwidth of Tier 2, W_2, increases, Tier-2 BSs can provide a higher service rate to the associated users. The optimal $\widetilde{B}_2^*$ should thus become larger so as to encourage more users to be associated with Tier-2 BSs. Moreover, it can be observed from Fig. 3.4(a) that as W_2/W increases, the optimal normalized biasing factor $\widetilde{B}_2^*$ becomes insensitive to the mean packet arrival rate of each user γ. The minimum lower bound of the network mean queuing delay $\bar{D}^*$, on the other hand, decreases as W_2/W increases, as Fig. 3.4(b) demonstrates.

While minimizing the network mean queuing delay is desirable for real-time traffic, the SIR coverage is an important performance metric to support non-real-time traffic for service providers. According to (2.13), the network SIR coverage S can be written as

$$S = \sum_{k=1}^{K} A_k \cdot \mathrm{P}\left[\mathrm{SIR}_k > \tau\right] = \sum_{k=1}^{K} \frac{A_k}{A_k \rho_k Z\left(\tau, \alpha, 1\right) + 1}. \tag{3.36}$$

Fig. 3.5(a) demonstrates how the network SIR coverage S varies with the normalized

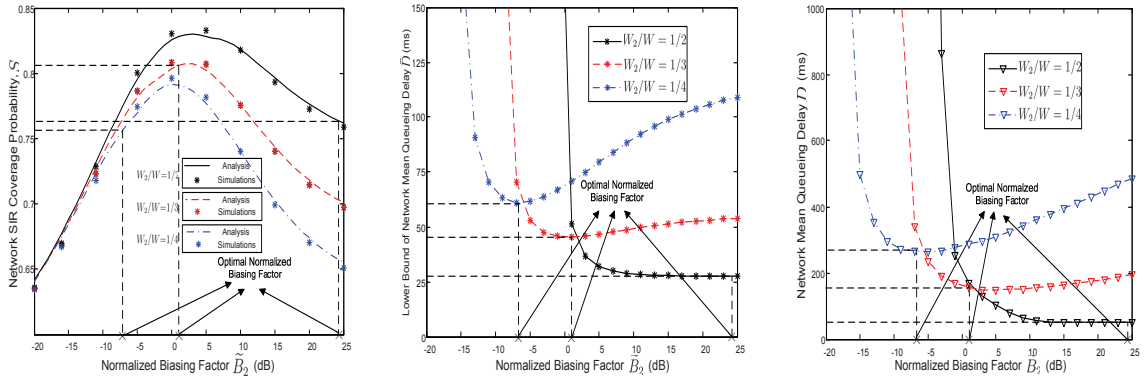


(a) (b)

Figure 3.4: Optimal normalized biasing factor $\widetilde{B_2^*}$ and the minimum lower bound of the network mean queuing delay $\bar{D}^*$ versus the bandwidth ratio of Tier 2 W_2/W with various values of the mean arrival bit rate per area $\gamma\lambda_u L$. W=12MHz and $\tau = 1$. (a) Optimal normalized biasing factor $\widetilde{B_2^*}$. (b) Minimum lower bound of the network mean queuing delay $\bar{D}^*$.

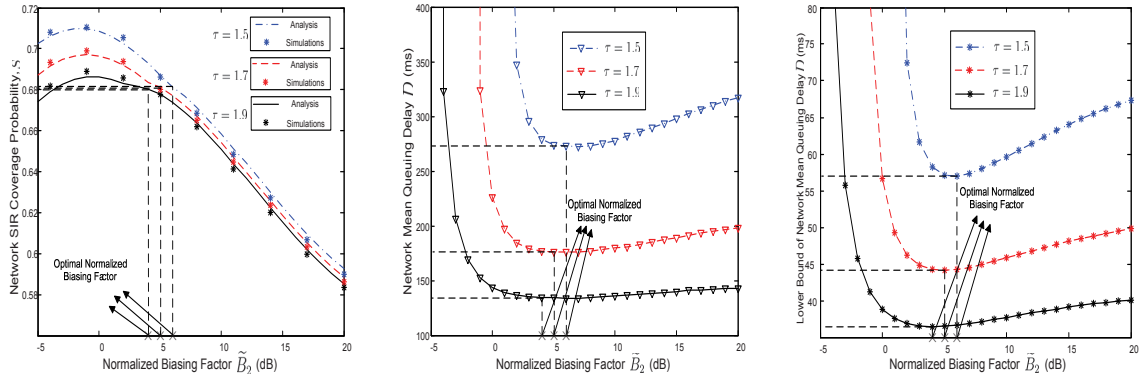
biasing factor $\widetilde{B}_2$ with various values of the bandwidth ratio W_2/W. It can be observed from Fig. 3.5(a) that there exists an optimal normalized biasing factor with which the network SIR coverage is maximized. Intuitively, when $\widetilde{B}_2$ is too large, a large fraction of users that are originally associated with Tier-1 BSs are offloaded to Tier-2 BSs. As these users are close to the interfering Tier-1 BSs and have long distances to their associated Tier-2 BSs, they have very poor channel conditions, which leads to a low SIR coverage of the network. Similarly, when $\widetilde{B}_2$ is too small, the network SIR coverage also deteriorates. In addition, it can be seen from Fig. 3.5(a) that the optimal normalized biasing to maximize S is insensitive to the bandwidth allocation. In the meanwhile, the optimal normalized biasing factor $\widetilde{B_2^*}$ to minimize $\bar{D}$ increases as W_2/W increases, as illustrated in Fig. 3.5(c) indicating a tradeoff between the network mean queuing delay and the network SIR coverage. For example, if $W_2/W = 1/2$, the optimal normalized biasing factor is obtained as $\widetilde{B_2^*} = 24$dB, with which the network SIR coverage greatly deteriorates. In this case, the service providers should properly tune the biasing factor in HetNets such that a desired point on the tradeoff curve can be achieved to balance the performances of real-time traffic and non-real-time traffic.

As the SIR threshold τ critically determines the network mean queuing delay and the network SIR coverage, Fig. 3.6 further demonstrates the impact of the SIR threshold τ on these two performance metrics. It can be observed from Fig. 3.6 that for a given normalized biasing factor $\widetilde{B}_2$, the network SIR coverage S decreases as the SIR threshold τ increases. In the meanwhile, both the network mean queuing delay D and its lower bound $\bar{D}$ decrease as τ increases. Intuitively, with a higher SIR threshold τ, the mean aggregate packet arrival rate of each BS becomes lower while the service rate becomes higher, leading to a better queuing performance. In addition, it is illustrated in Fig. 3.6(a) that the optimal normalized biasing factor to maximize the network SIR coverage S is insensitive to the SIR threshold τ, while the optimal normalized biasing factor $\widetilde{B_2^*}$ to



(a) (b) (c)

Figure 3.5: The network SIR coverage and the network mean queuing delay performance with various bandwidth ratios of Tier 2 W_2/W. $\gamma\lambda_u L = 18\text{Mbps/km}^2$, $W = 12\text{MHz}$, and $\tau = 1$. (a) Network SIR coverage S. (b) Network mean queuing delay D. (c) Lower bound of the network mean queuing delay $\bar{D}$.



(a) (b) (c)

Figure 3.6: The network SIR coverage and the network queuing delay performance with various values of the SIR threshold τ. $W_1 = 8\text{MHz}$, $W_2 = 4\text{MHz}$, and $\gamma\lambda_u L = 38\text{Mbps/km}^2$, (a) Network SIR coverage S. (b) Network mean queuing delay D. (c) Lower bound of the network mean queuing delay $\bar{D}$.

minimize $\bar{D}$ increases as τ decreases, as Fig. 3.6(c) demonstrates. Intuitively, although the service rates of both macro and micro BSs become lower with a smaller τ, macro BSs are more likely to become overloaded as their deployment density is much lower than that of micro BSs. The optimal normalized biasing factor $\widetilde{B_2^*}$ should thus become larger to undertake the load pressure from macro BSs. By comparing Fig. 3.6(a) with Fig. 3.6(b) and Fig. 3.6(c), it can be found that with a smaller SIR threshold τ, the deterioration of the network mean queuing delay D becomes much more severe if the normalized biasing factor is optimally tuned to maximize the network SIR coverage S, indicating a more significant tradeoff between the network SIR coverage and the network mean queuing delay.

3.5 Conclusion

In this chapter we have studied how to optimally tune the biasing factor of each tier in HetNets in order to minimize a lower bound of the network mean queuing delay. It is shown that the network queuing performance can be significantly improved when the biasing factor of each tier is optimally tuned. The characterization of the optimal biasing factor provides guidance for real-time service provisioning in HetNets. The case study of a 2-Tier HetNet further illustrates that the network mean queuing delay and the network SIR coverage might not be optimized simultaneously by tuning the biasing factor, indicating a performance tradeoff between real-time and non-real-time services.

It is worth mentioning that it is assumed that one BS will serve a user with a constant rate if its SIR exceeds a threshold. In practice, nevertheless, the service rate could depend on the channel conditions. In this case, as the biasing factor of one tier decreases, the mean service rate of this tier increases as the users located at the edge of the cells are offloaded. The queuing performance of this tier can thus be improved due to a lower mean

aggregate packet arrival rate and a higher mean service rate. Therefore, there would exist an optimal biasing factor for each tier such that the traffic load is balanced across tiers and the network mean queuing delay is minimized. On the other hand, if the biasing factor of one tier is too large, the SIR coverage of this tier degrades, which would drag down the network SIR coverage. Therefore, the network mean queuing delay may be optimized at the cost of the network SIR coverage.

Chapter 4

Queue-Aware Optimal Bandwidth Allocation in HetNets

In this chapter, we will study how to improve the network performance in terms of the energy efficiency and SIR coverage by properly allocating the spectrum resources to BSs of each tier in HetNets. By considering queuing in each BS, optimization problems to minimize the network average power consumption and to maximize the SIR coverage are formulated, which are shown to be convex and concave with respect to bandwidth allocation, respectively. When the mean packet arrival rate of each user is small, closed-form solutions to the optimization problems are obtained. Simulation results of a 2-Tier HetNet demonstrate that the network average power consumption and the SIR coverage can be significantly improved by the optimal spectrum allocation. A tradeoff between energy efficiency and SIR coverage is further revealed, which provides insights regarding the interplay of these two performance metrics.

4.1 Introduction

Besides the effort to optimize the delay performance in Chapter 3, energy efficiency becomes a crucial design factor for HetNets due the the high power consumption of the massive number of small-scale BSs. As the traffic activity varies distinctively during different time periods, an efficient approach is to put cells in a low energy mode during periods of low activity load [115, 116]. Hence, a lot of previous literatures [100, 101, 103, 104, 117] focused on this issue and improved the energy efficiency by introducing sleep-mode techniques where some BSs are selectively switched off according to the traffic load. As the aforementioned studies [100, 101, 103, 104, 117] assumed that one BS is always transmitting packets, the network energy consumption performance is thus only determined by the BS deployment intensity. As a result, the goal of the these studies is to determine the smallest set of active BSs to reduce energy consumption. With the consideration of queuing, nevertheless, there exists a significant gap of the power consumption of each individual BS between busy state and idle state. Therefore, the network energy consumption performance is also related to the queuing status of the BSs.

Recall that the average traffic intensity of each tier, i.e., the average BS busy probability of each tier, was derived in Chapter 2. With orthogonal spectrum allocation across tiers, the service rate of one tier would be higher with a wider bandwidth. Since one BS consumes less energy in the idle state, its average power consumption will be lower. Therefore, the bandwidth should be carefully allocated across tiers to achieve a high energy efficiency. There have been a great deal of effort in the previous literatures [68,118–122] towards spectrum allocation strategy. The authors in [119] proposed a bilateral bargaining algorithm to split the spectrum resource for a macro and micro BS link to maximize their achievable rates. To optimize the average downlink user data rate, Bao *et al.* [121, 122] proposed a structured spectrum allocation and user association scheme and showed that BSs of a tier with higher deployment density should have higher priority in spectrum

allocation. Sadr *et al.* [120] focused on a multi-tier HetNet and formulated a rate coverage maximization problem by properly allocating spectrum to each tier. Similarly, Lin *et al.* [68] obtained the optimal bandwidth allocation by maximizing the logarithm of the mean user rate. It was found in [68, 120] that the optimal proportion of spectrum allocated to a tier equals the proportion of users associated with that tier.

As it was assumed in [68, 118–122] that the BSs are always busy, energy efficiency becomes insensitive to the bandwidth allocation. Therefore, they only focused on improving spectrum efficiency. Hence, by taking queuing into account in this chapter, we will study the impact of the bandwidth allocation on the network energy efficiency and SIR coverage to find the optimal bandwidth allocation strategy. The contributions of this chapter are summarized as follows.

- Based on the derived average traffic intensity under the assumption of spectrum partitioning, a network average power consumption minimization problem as well as a network SIR coverage maximization problem with respect to the bandwidth allocation are formulated, which are shown to be convex and concave respectively.

- By using the approximation of the average traffic intensity of each tier, explicit solutions to the optimization problems are further derived.

- Simulation results demonstrate that both the network average power consumption and SIR coverage can be remarkably improved by properly allocating the bandwidth to each tier. In addition, it is further revealed that the network average power consumption and SIR coverage cannot be optimized simultaneously, which indicates a tradeoff between energy and SIR coverage.

The rest of this chapter is organized as follows. The system model is presented in Section 4.2. A network average power consumption minimization problem is studied in Section 4.3. A network SIR coverage maximization problem is examined in Section 4.4.

Simulation results are demonstrated in Section 4.5. Conclusion remarks are drawn in Section 4.6.

4.2 System Model

Consider a K-Tier heterogeneous network where BSs in the k^{th} tier, $k \in \{1, ..., K\}$, are deployed according to an independent homogenous PPP Φ_k with an intensity λ_k, and users form another independent homogenous PPP Φ_{u} with an intensity of λ_{u}. The instantaneous received power of a typical user from a BS in the k^{th} tier is given by $P_k g_k x_k^{-\alpha_k}$ where P_k is the transmission power of a BS in the k^{th} tier; g_k denotes the small-scale fading coefficient, which follows an i.i.d. exponential distribution of unit mean; and α_k is the path-loss coefficient, which is assumed to be identical across different tiers, i.e., $\alpha_k = \alpha$, $\forall k$. Each user associates with the BS with the largest average reference signal receiving power (RSRP). The set of BSs belonging to one tier operate in their own frequency band with the bandwidth W_k and hence do not interfere with the BSs of other tiers. Denote the total bandwidth as W. We then have $\sum_{k=1}^{K} W_k = W$.

Similar to the queuing model defined in Section 2.1.1, we assume that the packet request of each mobile user in the downlink follows an independent process with the mean arrival rate γ, and forms a queue in its associated BS. The packet length is exponentially distributed with the mean length L. Each BS transmits packets in a first-in-first-serve (FIFS) fashion. A signal-to-interference ratio (SIR) threshold τ is assumed, with which a BS will serve a user if the SIR exceeds τ, and drop the request otherwise. Therefore, by substituting $R_k = W_k \log_2(1+\tau)$ into (2.20), the average traffic intensity $\bar{\rho}_k$ can be written as

$$\bar{\rho}_k = \frac{-\lambda_k + \sqrt{\lambda_k^2 + \frac{4\gamma\lambda_{\text{u}} L \lambda_k Z A_k^2}{W_k \log_2(1+\tau)}}}{2A_k\lambda_k Z}, \tag{4.1}$$

where $Z = \tau^{\frac{2}{\alpha}} \int_{\tau^{-\frac{2}{\alpha}}}^{\infty} du / \left(1 + u^{\frac{\alpha}{2}}\right)$ and A_k denotes the association probability to a Tier-k

BSs, which can be found in (2.10). It is clear from (4.1) that the average traffic intensity of Tier k, $\bar{\rho}_k$, decreases as its allocated bandwidth W_k increases. As BSs consume more energy in the busy state than in the idle state, the energy consumption of the whole network can be optimized by a proper bandwidth allocation to balance the traffic load across different tiers.

4.3 Network Average Power Consumption Minimization

In this section, we will minimize the network average power consumption by optimally allocating the spectrum resources to each tier. Section 4.3.1 will first demonstrate the convexity of the optimization problem. Based on an approximation of the average traffic intensity, explicit solution of the problem is then obtained in Section 4.3.2.

4.3.1 Problem Formulation

According to [95], the power consumption of a Tier-k BS in the busy state can be written as

$$P_{k,\mathrm{T}} = P_{k,s} + \Delta_k P_k, \tag{4.2}$$

where $P_{k,s}$ is the power consumption of its signal processing and battery leakage, and Δ_k is the coefficient to reflect BS's cooling and feeder loss. On the other hand, the power consumption of a Tier-k BS in the idle state is given by

$$P_{k,\mathrm{I}} = \eta_k P_{k,\mathrm{T}}. \tag{4.3}$$

where $\eta_k < 1$ is the idle power coefficient, i.e., the BS power consumption ratio between the idle state and the busy state. The average power consumption of a BS in the k^{th} tier

can then be written as

$$P_{k,\mathrm{av}} = P_{k,\mathrm{T}}\bar{\rho}_k + P_{k,\mathrm{I}}\left(1-\bar{\rho}_k\right) = \left(1-\eta_k\right)P_{k,\mathrm{T}}\bar{\rho}_k + \eta_k P_{k,\mathrm{T}}. \tag{4.4}$$

By combining (4.1) and (4.4), the network average power consumption is given by

$$P = \sum_{k=1}^{K} \lambda_k P_{k,\mathrm{av}} = \sum_{k=1}^{K} \lambda_k \left(1-\eta_k\right) P_{k,\mathrm{T}}\bar{\rho}_k + \lambda_k \eta_k P_{k,\mathrm{T}}$$

$$= P_{k,\mathrm{T}} \sum_{k=1}^{K} \frac{\left(1-\eta_k\right)\left(-\lambda_k + \sqrt{\lambda_k^2 + \frac{4\gamma\lambda_{\mathrm{u}} L \lambda_k Z A_k^2}{W_k \log_2(1+\tau)}}\right)}{2A_k Z} + \lambda_k \eta_k. \tag{4.5}$$

It can be seen from (4.5) that the network average power consumption P is critically determined by the bandwidth allocation of each tier.

To minimize the network average power consumption, we have

$$\min_{\{W_k\}_{\forall k \in \{1,\ldots,K\}}} P_{k,\mathrm{T}} \sum_{k=1}^{K} \frac{\left(1-\eta_k\right)\left(-\lambda_k + \sqrt{\lambda_k^2 + \frac{4\gamma\lambda_{\mathrm{u}} L \lambda_k Z A_k^2}{W_k \log_2(1+\tau)}}\right)}{2A_k Z} + \lambda_k \eta_k, \tag{4.6a}$$

$$\text{s.t.} \sum\nolimits_{k=1}^{K} W_k = W, \tag{4.6b}$$

$$\rho_k < 1,\ k \in \{1, ..., K\}. \tag{4.6c}$$

The constraint (4.6b) comes from the fact that the bandwidth of each tier sums up to the total bandwidth W. Recall in Section 2.1.1 that in this thesis we focus on the condition where the traffic intensity $\rho_{k,i} \leq 1$, i.e., $\rho_{k,i}$ equals the busy probability of the BS. The constraint (4.6c) thus guarantees that the average traffic intensity equals the average BS busy probability for each tier, which can be converted into the bandwidth constraint

$$W_k > \frac{\gamma\lambda_{\mathrm{u}} L A_k}{\left(1 + A_k Z\right)\lambda_k \log_2\left(1+\tau\right)},\ k \in \{1, ..., K\}, \tag{4.7}$$

according to (4.1).

Before solving the optimization problem, first note that (4.6) does not have a feasible region $\mathbf{R}$ if

$$\sum_{k=1}^{K} \frac{\gamma \lambda_{\mathrm{u}} L A_k}{\left(1 + A_k \bar{\rho}_k Z\right) \bar{\rho}_k \lambda_k \log_2\left(1+\tau\right)} > W \tag{4.8}$$

according to (4.6b) and (4.7). Intuitively, one tier should be allocated a large bandwidth to improve the service rate to satisfy $\bar{\rho}_k < 1$. In this case, the total bandwidth W might not be wide enough to meet the requirement of $\bar{\rho}_k < 1$.

If (4.8) doest not hold, we could minimize the network average power consumption per area, P, by properly allocating the bandwidth to each tier. The following lemma proves the convexity of the optimization problem (4.6).

Lemma 4.1. *The network average power consumption per area P is convex with respect to the bandwidth allocation $\{W_k\}_{\forall k}$ under the constraints of* (4.6b) *and* (4.6c).

Proof. According to (4.1), the first and second order derivative of $\bar{\rho}_k$ with respect to W_k can be obtained as

$$\frac{d\bar{\rho}_k}{dW_k} = -\frac{\gamma \lambda_{\mathrm{u}} L A_k}{\beta^{\frac{1}{2}} \log_2\left(1+\tau\right) W_k^2} > 0, \tag{4.9}$$

and

$$\begin{aligned}\frac{d^2\bar{\rho}_k}{{dW_k}^2} &= \frac{\gamma \lambda_{\mathrm{u}} L A_k}{\log_2\left(1+\tau\right)} \cdot \left[2\beta^{-\frac{1}{2}} W_k^{-3} - \frac{2\beta^{-\frac{3}{2}} W_k^{-4} \gamma \lambda_{\mathrm{u}} L \lambda_k A_k^2 Z}{\log_2\left(1+\tau\right)}\right] \\ &= \frac{\gamma \lambda_{\mathrm{u}} L A_k}{2\log_2\left(1+\tau\right) \beta^{\frac{3}{2}} W_k^4} \left[\lambda_k^2 W_k + \frac{16\gamma \lambda_{\mathrm{u}} L \lambda_k A_k^2 Z}{\log_2\left(1+\tau\right)} - \frac{4\gamma \lambda_{\mathrm{u}} L \lambda_k A_k^2 Z}{\log_2\left(1+\tau\right)}\right] > 0, \end{aligned} \tag{4.10}$$

respectively, where

$$\beta = \lambda_k^2 + \frac{4\gamma \lambda_{\mathrm{u}} L \lambda_k A_k^2 Z}{\log_2\left(1+\tau\right) W_k}. \tag{4.11}$$

Therefore, the average traffic intensity $\bar{\rho}_k$ is convex with respect to the bandwidth allocation of each tier. As the network average power consumption P is a linear function of $\bar{\rho}_k$, P is convex with respect to W_k. Finally, the constraint (4.6c) is equivalent to the constraint (4.7). As both (4.6b) and (4.6c) are linear constraints, the network average power consumption minimization problem (4.6) is convex. □

Intuitively, the spectrum allocated to one tier cannot be too small or too large. If it is too small, BSs of this tier would be more likely to be busy, and thus consume more power. Otherwise, BSs of other tiers would have a higher busy probability such that the network average power consumption would also deteriorate.

4.3.2 Explicit Solution

According to Lemma 4.1, the optimal bandwidth allocated to each tier $\{W_k^{*,P}\}_{\forall_k}$ can be obtained by a numerical method such as gradient decent. However, to obtain the explicit solution of the network average power minimization problem (4.6), we adopt a similar approximation of the average traffic intensity $\bar{\rho}_k$ by (3.7) in Section 3.3.1. In this case, the network average power consumption P can be further written as

$$P{=}\gamma\lambda_{\mathrm{u}}L\sum\nolimits_{k=1}^{K}\frac{(1-\eta_k)\,P_{k,\mathrm{T}}A_k}{W_k\log_2(1+\tau)}+\sum\nolimits_{k=1}^{K}\lambda_k\eta_k P_{k,\mathrm{T}}, \tag{4.12}$$

and the constraint (4.6c) is equivalent to the bandwidth constraint

$$W_k>\frac{\gamma\lambda_{\mathrm{u}}LA_k}{\lambda_k\log_2(1+\tau)},\ k\in\{1,...,K\}, \tag{4.13}$$

according to (3.7). The following lemma gives an explicit solution of the bandwidth allocation $\{W_k^{*,P}\}_{\forall k}$ to the optimization problem (4.6).

Lemma 4.2. *When the mean packet arrival rate of each user satisfies*

$$\gamma<\min_{\forall k}\left\{\frac{W\log_2(1+\tau)\sqrt{\lambda_k^3P_k^{\frac{2}{\alpha}}(1-\eta_k)\,P_{k,T}}}{\lambda_{\mathrm{u}}LA_k\sum\limits_{i=1}^{K}\sqrt{\lambda_iP_i^{\frac{2}{\alpha}}(1-\eta_i)\,P_{i,T}}}\right\}, \tag{4.14}$$

the optimal bandwidth allocation $\{W_k^{*,P}\}_{\forall k}$ can be written as

$$W_k^{*,P}=\frac{\sqrt{\lambda_kP_k^{\frac{2}{\alpha}}(1-\eta_k)\,P_{k,T}}}{\sum_{i=1}^{K}\sqrt{\lambda_iP_i^{\frac{2}{\alpha}}(1-\eta_i)\,P_{i,T}}}W,\ k\in\{1,...K\}. \tag{4.15}$$

Proof. According to (3.7), the second-order derivative of $\bar{\rho}_k$ with respect to W_k can be obtained as

$$\frac{d^2\bar{\rho}_k}{dW_k{}^2} = \frac{2\gamma\lambda_{\mathrm{u}}LA_k}{\lambda_k \log_2(1+\tau) W_k^3} > 0. \tag{4.16}$$

As the network average power consumption P is a linear function of $\bar{\rho}_k$, (4.12) is convex with respect to $\{W_k\}_{\forall k}$. Therefore, by substituting W_K with $W - \sum_{k=1}^{K-1} W_k$ in (4.12), and taking the partial derivative of P with respect to W_k, we have

$$\frac{\partial P}{\partial W_k} = -\frac{\gamma\lambda_{\mathrm{u}}LA_k\Delta_k P_k}{W_k^2 \log_2(1+\tau)} + \frac{\gamma\lambda_{\mathrm{u}}LA_K\Delta_K P_K}{\left(W - \sum\limits_{k=1}^{K-1} W_k\right)^2 \log_2(1+\tau)},\ k = 1, ..., K-1. \tag{4.17}$$

The optimal bandwidth allocation $\{W_k^{*,P}\}_{\forall k}$ can be obtained by solving the set of equations $\frac{\partial P}{\partial W_k}$=0, $k = 1, ..., K-1$ as

$$W_k^{*,P} = \frac{\sqrt{\lambda_k P_k^{\frac{2}{\alpha}} (1-\eta_k) P_{k,T}}}{\sum\limits_{i=1}^{K} \sqrt{\lambda_i P_i^{\frac{2}{\alpha}} (1-\eta_i) P_{i,T}}} W,\ k \in \{1, ...K\}. \tag{4.18}$$

Finally, by combining (4.13) with (4.18), we have

$$W_k^{*,P} = \frac{\sqrt{\lambda_k P_k^{\frac{2}{\alpha}} (1-\eta_k) P_{k,T}}}{\sum\limits_{i=1}^{K} \sqrt{\lambda_i P_i^{\frac{2}{\alpha}} (1-\eta_i) P_{i,T}}} W > \frac{\gamma\lambda_{\mathrm{u}}LA_k}{\lambda_k \log_2(1+\tau)}, \tag{4.19}$$

which leads to the constraint of the mean packet arrival rate in Lemma 4.2 as

$$\gamma < \min_{\forall k} \left\{ \frac{W \log_2(1+\tau) \sqrt{\lambda_k^3 P_k^{\frac{2}{\alpha}} (1-\eta_k) P_{k,T}}}{\lambda_{\mathrm{u}} LA_k \sum\limits_{i=1}^{K} \sqrt{\lambda_i P_i^{\frac{2}{\alpha}} (1-\eta_i) P_{i,T}}} \right\}. \tag{4.20}$$

□

With the optimal bandwidth allocation, when γ reaches the upper bound, there must exist a certain tier, Tier-k for instance, such that its average traffic intensity $\bar{\rho}_k$ approaches 1. In this case, γ is not small compared to the service rate of the queue in each Tier-k

BS. On the other hand, since (3.7) becomes closer to (2.20) as γ decreases, the optimal bandwidth allocation obtained in (4.15) is more accurate with a smaller value of γ. As the network average power consumption in (4.12) is a linear function of the mean packet arrival rate γ, the optimal bandwidth allocation $W_k^{*,P}$ does not depend on γ. In addition, it can be clearly seen from (4.15) that $W_k^{*,P}$ increases as the transmission power P_k or the BS deployment density λ_k increases. Intuitively, with a larger P_k or a higher λ_k, users are more likely to associate with a Tier-k BS. A larger bandwidth can effectively reduce the power consumption of this tier.

4.4 Network SIR Coverage maximization

Besides energy efficiency, the SIR coverage is another important performance metric which characterizes spectrum efficiency [119–122]. In this section, we will further derive the optimal bandwidth allocation to maximize the network SIR coverage. Section 4.4.1 will first demonstrate the convexity of the optimization problem. Based on an approximation of the average traffic intensity, explicit solution of the problem is then obtained in Section 4.4.2.

4.4.1 Problem Formulation

Recall that the SIR coverage of Tier-k BSs has been derived as

$$S_k = \mathrm{P}\left[\mathrm{SIR}_k > \tau\right] = \frac{1}{A_k \rho_k Z + 1}. \tag{4.21}$$

according to (2.13). By combining (4.1) and (4.21), the network SIR coverage S is given by

$$S = \sum_{k=1}^{K} \frac{A_k}{A_k \rho_k Z + 1} = \sum_{k=1}^{K} \frac{2\lambda_k A_k}{\lambda_k + \sqrt{\lambda_k^2 + \frac{4\gamma\lambda_{\mathrm{u}} L \lambda_k Z A_k^2}{W_k \log_2(1+\tau)}}}. \tag{4.22}$$

To maximize the network SIR coverage, we have

$$\max_{\{W_k\}_{\forall k\in\{1,\ldots,K\}}} = \sum_{k=1}^{K} \frac{2\lambda_k A_k}{\lambda_k + \sqrt{\lambda_k^2 + \frac{4\gamma\lambda_u L\lambda_k Z A_k^2}{W_k \log_2(1+\tau)}}}, \tag{4.23a}$$

$$\text{s.t.} \sum\nolimits_{k=1}^{K} W_k = W, \tag{4.23b}$$

$$\rho_k < 1,\ k \in \{1, ..., K\}. \tag{4.23c}$$

Similarly, (4.23) does not have a feasible region $\mathbf{R}$ if (4.8) holds. The following lemma proves the concavity of (4.23).

Lemma 4.3. *The network SIR coverage S is concave with respect to the bandwidth allocation $\{W_k\}_{\forall k}$ under the constraint of* (4.23b) *and* (4.23c)*.*

Proof. According to (4.21), the second-order derivative of Tier-k SIR coverage S_k with respect to W_k can be obtained as

$$\frac{d^2 S_k}{{dW_k}^2} = -A_k^2 Z \left[\frac{-2A_k Z}{\left(A_k \bar{\rho}_k Z + 1\right)^3} \cdot \frac{d\bar{\rho}_k}{dW_k} + \frac{1}{\left(A_k \bar{\rho}_k Z + 1\right)^2} \cdot \frac{d^2 \bar{\rho}_k}{{dW_k}^2} \right]. \tag{4.24}$$

Recall that we have $\frac{d\bar{\rho}_k}{dW_k}<0$ and $\frac{d^2\bar{\rho}_k}{{dW_k}^2}>0$ according to (4.16) and (4.17). It can then be concluded that $\frac{\partial^2 S_k}{\partial {W_k}^2}<0$. Since the network SIR coverage can be written as $S = \sum\limits_{k=1}^{K} A_k S_k$, which is a linear function of the SIR coverage of Tier k, the network SIR coverage S is concave with respect to $\{W_k\}_{\forall k}$. Finally, as the constraints (4.23b) and (4.23c) are linear, the network SIR coverage maximization problem is concave. □

Intuitively, when the bandwidth allocated to a certain tier is small, BSs of this tier would cause severe interference to the associated users due to a high busy probability. Otherwise, the interference level in other tier would be high. In both cases, the network SIR coverage would deteriorate.

4.4.2 Explicit Solution

Lemma 4.3 guarantees that the optimal bandwidth allocation $\{W_k^{*,S}\}_{\forall k}$ to maximize the network SIR coverage S can be obtained by numerical approaches such as the gradient decent method. Similar to the derivations of $\{W_k^{*,P}\}_{\forall k}$ in Section 4.3.2, we use (3.7) to approximate the average traffic intensity $\bar{\rho}_k$. The network SIR coverage can then be rewritten as

$$S = \sum\nolimits_{k=1}^{K} \frac{\lambda_k W_k A_k \log_2(1+\tau)}{Z\gamma\lambda_{\mathrm{u}} L A_k^2 + \lambda_k W_k \log_2(1+\tau)}, \tag{4.25}$$

and the constraint (4.23c) can be further written as

$$W_k > \frac{\gamma\lambda_{\mathrm{u}} L A_k}{\lambda_k \log_2(1+\tau)},\ k \in \{1, ..., K\}, \tag{4.26}$$

by combining (3.7) and (4.22). The following lemma presents an explicit expression of the optimal bandwidth allocation $\{W_k^{*,S}\}_{\forall k}$.

Lemma 4.4. *When the mean packet arrival rate of each user satisfies*

$$\gamma < \min_{\forall k} \left\{ \frac{W\log_2(1+\tau)\left(\sum_{j=1}^{K}\lambda_j P_j^{\frac{3}{\alpha}}\right)^2 \lambda_k^2 P_k^{\frac{3}{\alpha}}}{\lambda_{\mathrm{u}} L\left[Z\lambda_k^2 P_k^{\frac{3}{\alpha}} \sum_{l=1}^{K}\lambda_l P_l^{\frac{3}{\alpha}}\left(P_k^{\frac{1}{\alpha}} - P_l^{\frac{1}{\alpha}}\right) - A_k\left(\sum_{j=1}^{K}\lambda_j P_j^{\frac{3}{\alpha}}\right)^2 \sum_{j=1}^{K}\lambda_j P_j^{\frac{3}{\alpha}}\right]} \right\}, \tag{4.27}$$

the optimal bandwidth allocation $\{W_k^{*,S}\}_{\forall k}$ *can be written as*

$$W_k^{*,S} = \frac{\lambda_k P_k^{\frac{3}{\alpha}}}{\sum_{j=1}^{K}\lambda_j P_j^{\frac{3}{\alpha}}} \cdot \left[W - \frac{\gamma\lambda_{\mathrm{u}} L Z}{\log_2(1+\tau)} \cdot \frac{\sum_{l=1}^{K}\lambda_l P_l^{\frac{3}{\alpha}}\left(P_k^{\frac{1}{\alpha}} - P_l^{\frac{1}{\alpha}}\right)}{\left(\sum_{j=1}^{K}\lambda_j P_j^{\frac{2}{\alpha}}\right)^2}\right],\ k \in \{1, ..., K\}. \tag{4.28}$$

Proof. By substituting (3.7) into (4.21), the second-order derivative of S_k with respect to W_k can be obtained as

$$\frac{d^2 S_k}{dW_k{}^2} = -\frac{2\gamma\lambda_{\mathrm{u}} L Z A_k^3 \lambda_k^2 \log_2^2(1+\tau)}{\left[\gamma\lambda_{\mathrm{u}} L Z A_k^2 + \lambda_k W_k \log_2(1+\tau)\right]^3} < 0. \tag{4.29}$$

As the network SIR coverage S is a linear function of S_k, (4.25) is concave with respect to $\{W_k\}_{\forall k}$. Therefore, by substituting W_K with $W-\sum_{k=1}^{K-1}W_k$ in (4.25), and taking the partial derivative of S with respect to W_k, we have

$$\frac{\partial C}{\partial W_k}=\frac{\gamma\lambda_{\mathrm{u}}LZA_K^3\lambda_K\log_2(1+\tau)}{\left[\gamma\lambda_{\mathrm{u}}LZA_K^2+\lambda_K\left(W-\sum\limits_{j=1}^{K-1}W_j\right)\log_2(1+\tau)\right]^2} - \frac{\gamma\lambda_{\mathrm{u}}LZA_k^3\lambda_k\log_2(1+\tau)}{\left[\gamma\lambda_{\mathrm{u}}LZA_k^2+\lambda_kW_k\log_2(1+\tau)\right]^2},\ k=1,...,K-1. \quad (4.30)$$

The optimal bandwidth allocation $\{W_k^{*,S}\}_{\forall k}$ can be obtained by solving the set of equations $\frac{\partial S}{\partial W_k}$=0, $k=1,...,K-1$ as

$$W_k^{*,S}=\frac{\lambda_kP_k^{\frac{3}{\alpha}}}{\sum\limits_{j=1}^{K}\lambda_jP_j^{\frac{3}{\alpha}}}\cdot\left[W-\frac{\gamma\lambda_{\mathrm{u}}LZ}{\log_2(1+\tau)}\cdot\frac{\sum\limits_{l=1}^{K}\lambda_lP_l^{\frac{3}{\alpha}}\left(P_k^{\frac{1}{\alpha}}-P_l^{\frac{1}{\alpha}}\right)}{\left(\sum\limits_{j=1}^{K}\lambda_jP_j^{\frac{2}{\alpha}}\right)^2}\right],\ k\in\{1,...K\}. \quad (4.31)$$

Finally, by combining (4.26) with (4.31), we have

$$W_k^{*,S}=\frac{\lambda_kP_k^{\frac{3}{\alpha}}}{\sum\limits_{j=1}^{K}\lambda_jP_j^{\frac{3}{\alpha}}}\cdot\left[W-\frac{\gamma\lambda_{\mathrm{u}}LZ}{\log_2(1+\tau)}\cdot\frac{\sum\limits_{l=1}^{K}\lambda_lP_l^{\frac{3}{\alpha}}\left(P_k^{\frac{1}{\alpha}}-P_l^{\frac{1}{\alpha}}\right)}{\left(\sum\limits_{j=1}^{K}\lambda_jP_j^{\frac{2}{\alpha}}\right)^2}\right]>\frac{\gamma\lambda_{\mathrm{u}}LA_k}{\lambda_k\log_2(1+\tau)}>0, \quad (4.32)$$

which leads to the constraint of the mean packet arrival rate in Lemma 4.4 as

$$\gamma<\min_{\forall k}\left\{\frac{W\log_2(1+\tau)\left(\sum\limits_{j=1}^{K}\lambda_jP_j^{\frac{3}{\alpha}}\right)^2\lambda_k^2P_k^{\frac{3}{\alpha}}}{\lambda_{\mathrm{u}}L\left[Z\lambda_k^2P_k^{\frac{3}{\alpha}}\sum\limits_{l=1}^{K}\lambda_lP_l^{\frac{3}{\alpha}}\left(P_k^{\frac{1}{\alpha}}-P_l^{\frac{1}{\alpha}}\right)-A_k\left(\sum\limits_{j=1}^{K}\lambda_jP_j^{\frac{3}{\alpha}}\right)^2\sum\limits_{j=1}^{K}\lambda_jP_j^{\frac{3}{\alpha}}\right]}\right\}. \quad (4.33)$$

□

When γ is small, the optimal bandwidth allocated $W_k^{*,S}$ can be further written as

$$W_k^{*,S}\Big|_{\gamma\to0}=\frac{\lambda_kP_k^{\frac{3}{\alpha}}}{\sum_{j=1}^{K}\lambda_jP_j^{\frac{3}{\alpha}}}W, \quad (4.34)$$

Table 4.1: Simulation Parameters

Parameter	Value
User Density λ_u	10^{-3} m^{-2}
Tier-1 BS Density λ_1	10^{-5} m^{-2}
Tier-2 BS Density λ_2	$5*10^{-5}$ m^{-2}
Tier-1 Power Consumption Coefficient Δ_1	4.7
Tier-2 Power Consumption Coefficient Δ_2	2.6
Tier-1 Idle Power Coefficient η_1	0.1
Tier-2 Idle Power Coefficient η_2	0.1
Total Bandwidth W	12 MHz
Path Loss Coefficient α	4
SIR Threshold τ	1
Mean Packet Length L	0.001 Mb

which increases as the transmission power P_k or the BS intensity λ_k increases. Intuitively, with a larger P_k or λ_k, more users are associated to this tier, leading to a higher BS busy probability and a higher intra-tier interference level. By allocating more spectrum resources to this tier, the network SIR coverage can be improved.

4.5 Simulation Results

In this section, we will demonstrate simulation results to validate the analysis in Section 4.3 and Section 4.4. We consider a 2-Tier HetNet where the total available bandwidth is set to be fixed. Each tier occupies an orthogonal spectrum band. We perform Monte Carlo simulations over different topologies where BSs of each tier and users are drawn from independent PPPs with given intensities in each topology. Each point of the demonstrated

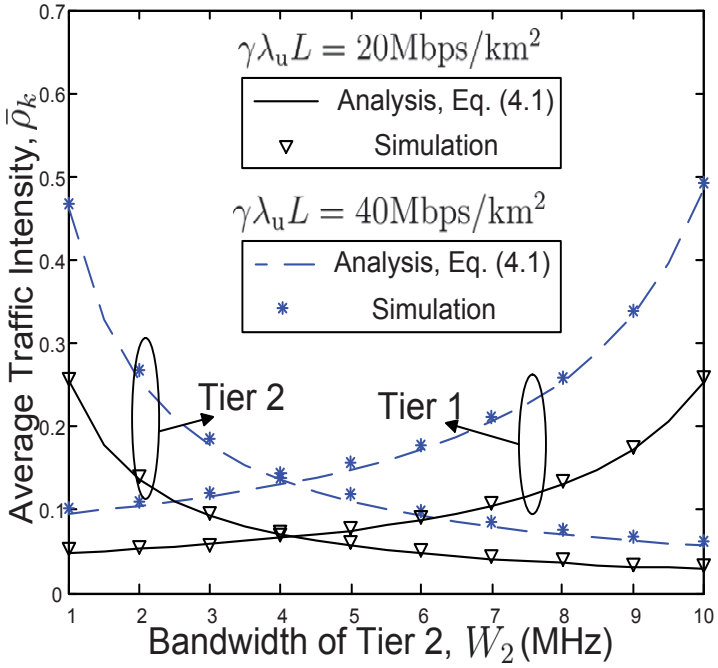


Figure 4.1: Average traffic intensity $\bar{\rho}_k$ of each tier versus the bandwidth of Tier 2, W_2. $P_1 = 20$W, $P_2 = 6$W.

results is obtained by averaging over 500 simulations on a time scale of 10^5 seconds. The parameters regarding BS energy consumption are set according to [95]. For simplicity, we assume the idle power coefficient $\eta_1 = \eta_2 = 0.1$. Table 4.1 summarizes the system parameters.

An explicit expression of the average traffic intensity with respect to the bandwidth allocation has been shown as (4.1) in Section 4.2. Fig. 4.1 illustrates how the average traffic intensities of the BSs of both tiers, i.e., $\bar{\rho}_1$ and $\bar{\rho}_2$, vary with the bandwidth of Tier-2, W_2, under various values of the mean bit arrival rate per area $\gamma\lambda_u L$. It can be observed from Fig. 4.1 that as the bandwidth W_2 increases, the average traffic intensity of Tier 1 increases, while that of Tier 2 decreases. Intuitively, with a higher bandwidth, the service rate of Tier 2 becomes larger, leading to a better queuing performance.

The network average power consumption has been characterized in Section 4.3.1 based on the derived average traffic intensity $\bar{\rho}_k$. Fig. 4.2 demonstrates how the network average

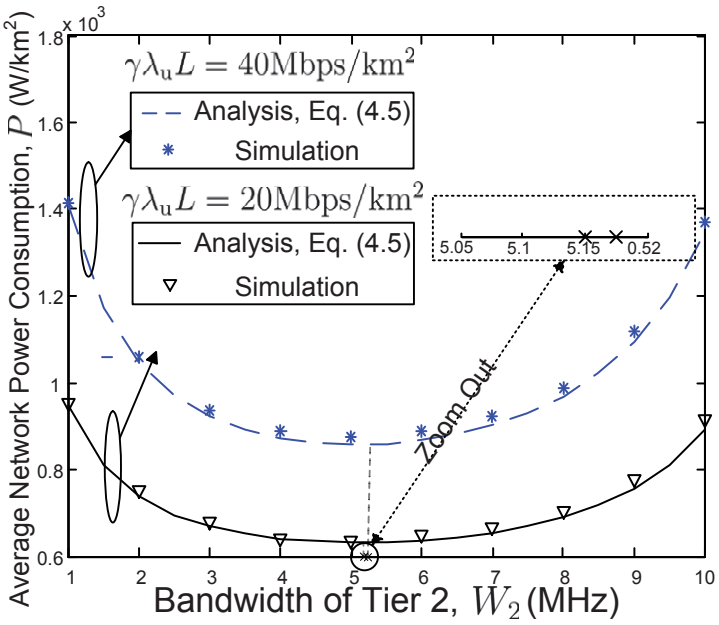


Figure 4.2: Network average power consumption P versus the bandwidth of Tier 2, W_2. $P_1 = 20$W, $P_2 = 6$W.

power consumption, P, varies with W_2 under various values of the mean bit arrival rate per area $\gamma\lambda_u L$. It can be observed that the network average power consumption P is sensitive to the bandwidth allocation. By optimally tuning the bandwidth allocation, the network average power consumption can be minimized. For instance, with $\gamma\lambda_u L = 40$ Mbps/km^2, the optimal bandwidth of Tier 2 can be obtained as $W_2^{*,P} = 5.15$ MHz, and the network average power consumption is as low as $0.86 * 10^3$ W/km^2.

On the other hand, the network SIR coverage has been derived as (4.22) in Section 4.4.2. Fig. 4.3 illustrates how the network SIR coverage, S, varies with the bandwidth of Tier 2, W_2. It can be observed from Fig. 4.3 that the network SIR coverage is concave with respect to the bandwidth of Tier 2. Intuitively, as W_2 increases, Tier-2 BSs are more likely to be idle, and the interference in Tier 2 becomes less severe. However, as W_2 becomes larger, Tier-1 BSs are more probable to be busy, indicating that the interference in Tier 1 degrades, leading to the deterioration of the whole network SIR coverage. By

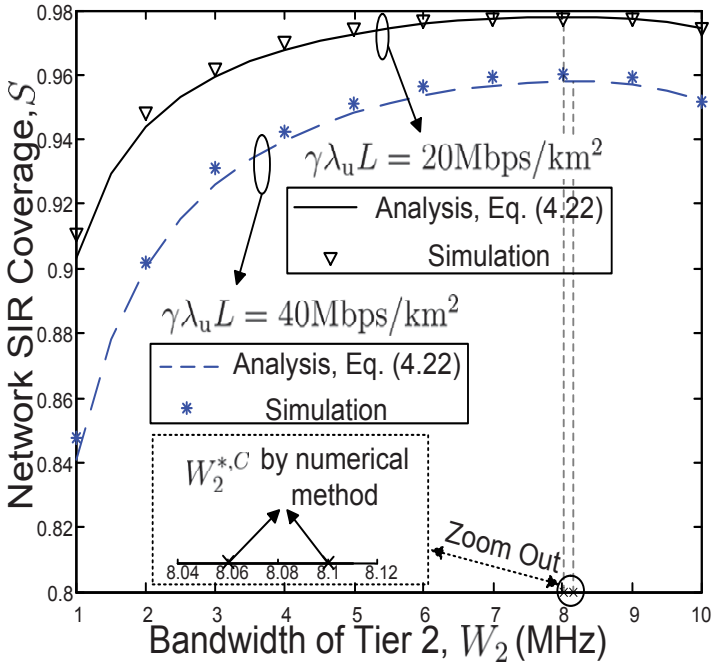


Figure 4.3: Network SIR coverage S versus the bandwidth of Tier 2, W_2. $P_1 = 20$W, $P_2 = 6$W.

comparing Fig. 4.2 with Fig. 4.3, it can be seen that the optimal bandwidth allocation to minimize the network average power consumption is not the same as that to maximize the network SIR coverage. Therefore, the network SIR coverage and the network power consumption could not be optimized simultaneously, indicating a tradeoff between these two performance metrics. In practice, the service providers should properly determine a desired point to balance between SIR coverage and energy efficiency.

Fig. 4.4 further demonstrates how the optimal bandwidth of Tier 2 to minimize the network average power consumption and to maximize the network SIR coverage, i.e., $W_2^{*,P}$ and $W_2^{*,S}$, vary with the BS density ratio, λ_2/λ_1, with various values of the BS transmission power of Tier 2, P_2. Besides the significant gap between $W_2^{*,P}$ and $W_2^{*,S}$ which is discussed in the previous paragraph, it can be observed from Fig. 4.4 that both $W_2^{*,P}$ and $W_2^{*,S}$ increase as the density of Tier-2 λ_2 or the transmission power of Tier 2 P_2 increases. Intuitively, with a larger P_2 or λ_2, the association probability of Tier 2

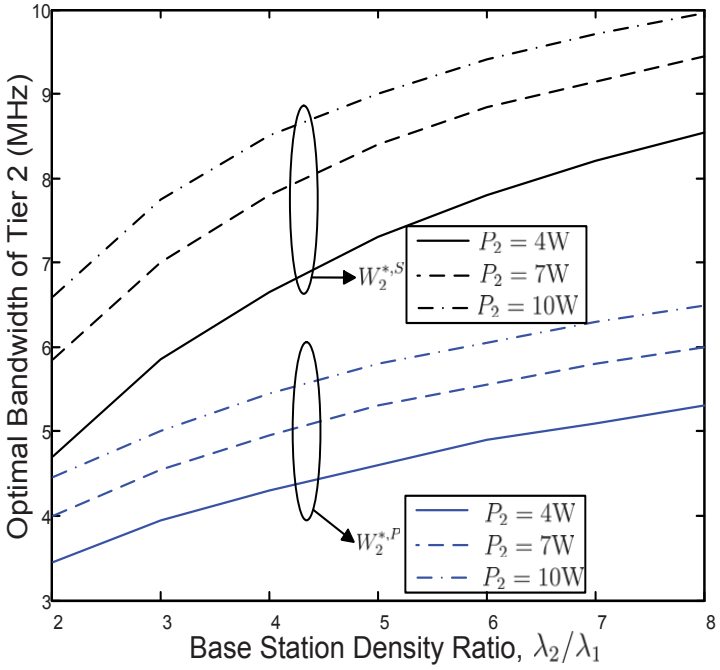


Figure 4.4: Optimal bandwidth of Tier 2, $W_2^{*,P}$ and $W_2^{*,S}$, to minimize the network average power consumption and maximize the network SIR coverage. $P_1 = 20$W and $\gamma\lambda_{\mathrm{u}}L = 40\mathrm{Mbps/km}^2$.

becomes larger. Tier 2 thus has a higher busy probability and intra-tier interference level, which imposes a significant impact on both the average power consumption P and the SIR coverage S of the network. Hence, more spectrum resources should be allocated to Tier 2 to improve the queuing performance of this tier.

To show the accuracy of the results in Lemma 4.2 and Lemma 4.4, Fig. 4.5 demonstrates the optimal bandwidth of Tier 2, i.e., $W_2^{*,P}$ and $W_2^{*,S}$, as well as the corresponding minimum network average power consumption P^* and maximum network SIR coverage S^*. Note that the range of the mean bit arrival rate per area in Fig. 4.5 is obtained as

$$\gamma\lambda_{\mathrm{u}}L < \min_{\forall k}\left\{\frac{W\log_2\left(1+\tau\right)\sqrt{\lambda_k^3 P_k^{\frac{2}{\alpha}}\left(1-\eta_k\right)P_{k,T}}}{A_k\sum\limits_{i=1}^{K}\sqrt{\lambda_i P_i^{\frac{2}{\alpha}}\left(1-\eta_i\right)P_{i,T}}}\right\} \tag{4.35a}$$

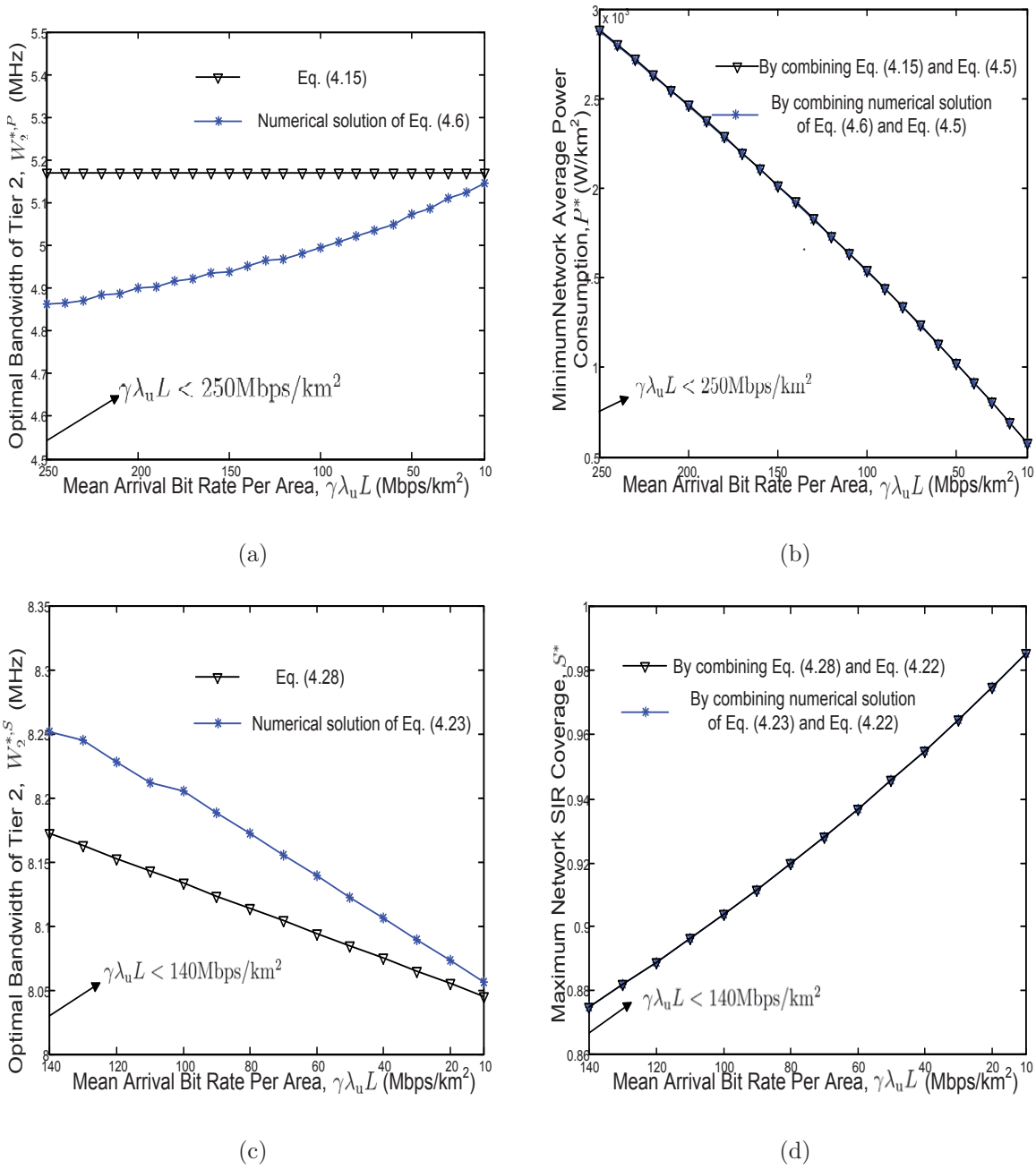


Figure 4.5: Optimal bandwidth of Tier 2 $W_2^{*,P}$ and $W_2^{*,S}$ as well as the corresponding minimum network average power consumption P^* and maximum network SIR coverage S^* versus the mean arrival bit rate per area $\gamma\lambda_u L$. (a) Optimal bandwidth of Tier 2, $W_2^{*,P}$. (b) Minimum network average power consumption P^*. (c) Optimal bandwidth of Tier 2, $W_2^{*,S}$. (d) Maximum network SIR coverage S^*.

and

$$\gamma\lambda_{\mathrm{u}}L < \min_{\forall k}\left\{\frac{W\log_2\left(1+\tau\right)\left(\sum\limits_{j=1}^{K}\lambda_j P_j^{\frac{3}{\alpha}}\right)^2\lambda_k^2 P_k^{\frac{3}{\alpha}}}{Z\lambda_k^2 P_k^{\frac{3}{\alpha}}\sum\limits_{l=1}^{K}\lambda_l P_l^{\frac{3}{\alpha}}\left(P_k^{\frac{1}{\alpha}}-P_l^{\frac{1}{\alpha}}\right)-A_k\left(\sum\limits_{j=1}^{K}\lambda_j P_j^{\frac{3}{\alpha}}\right)^2\sum\limits_{j=1}^{K}\lambda_j P_j^{\frac{3}{\alpha}}}\right\} \tag{4.35b}$$

according to (4.14) and (4.27), respectively. It can be observed from Fig. 4.5(a) and Fig. 4.5(c) that the there is a small gap between the two curves, and the gap diminishes as $\gamma\lambda_{\mathrm{u}}L$ decreases. Furthermore, it can be observed from Fig. 4.5(b) and Fig. 4.5(d) that the two curves of P^* and S^* almost overlap, which indicates that the bandwidth allocation according to (4.15) and (4.28) can achieve near-optimal performance.

4.6 Conclusion

This chapter studies how to optimize the network power consumption and the network SIR coverage by allocating spectrum resources to each tier in HetNets. By considering queues in each BS, optimization problems to minimize the network average power consumption and to maximize the network SIR coverage are formulated, which are shown to be convex and concave, respectively. Closed-form solutions are given by using an approximation of the average traffic intensity. Simulation results of a 2-Tier HetNet demonstrate that the network average power consumption and network SIR coverage cannot be optimized simultaneously, indicating a tradeoff between energy efficiency and SIR coverage.

Chapter 5

Queue-Aware Energy Efficient BS Density Optimization in HetNets

In this section, we will study the minimization of network average power consumption in a 2-Tier HetNet by optimally tuning the activation ratio of micro BSs under the QoS constraints of the network mean queuing delay and the network SIR coverage. With the assumption of universal frequency reuse (UFR), the average traffic intensity of each tier is characterized by a set of fixed-point equations, which can be solved by a proposed iterative method. By using the approximation that BSs of a tier have the same SIR coverage, the cumulative distribution function (CDF) of the traffic intensity of each tier is then obtained. On that basis, the network average power consumption per area as well as the constraints of the network mean queuing delay and the network SIR coverage are characterized. Numerical results demonstrate that if the idle power coefficient is below a certain threshold, the optimal activation ratio should equal the one to minimize the network average power consumption per area. Otherwise, the optimal activation ratio should be obtained according to the QoS constraints. It is further shown that UFR outperforms spectrum partitioning (SP) in terms of both energy efficiency and spectrum

efficiency in the considered scenario.

5.1 Introduction

We have shown in Chapter 4 that the network energy efficiency can be significantly improved by optimally allocating the spectrum resources to each tier in the consideration of queuing, the most direct and effective way to reduce the network power consumption is to control of the BS deployment density.

5.1.1 Energy Efficiency Optimization

Although a great deal of effort [103, 104, 106, 107, 123] has been made to find the optimal BS density in HetNets, they all assumed that BSs are transmitting packets all the time. Therefore, a fixed power consumption of each individual BS was considered, and the total network power consumption thus linearly increases as the BS deployment density increases. By taking queuing into account, nevertheless, one BS consumes less energy in the idle state than it does in the busy state. Tuning the BS deployment density thus have a significant impact on the the probability that one BS is in the busy state, which in turn affects its power consumption. Hence, this motivates us to characterize the network average power consumption by taking consideration of the average traffic intensity. Under the assumption of queuing, we consider a 2-Tier HetNet with universal frequency reuse (UFR) across both tiers in this chapter, in which case both the intra-tier interference and the inter-tier interference exist. According to the set of fixed-point equations of the average traffic intensity derived in Section 2.3.2, the existence and the uniqueness of the solutions are further proved in the considered scenario. An iterative method is then proposed to numerically obtain the average traffic intensity. On that basis, an optimization problem is formulated to minimize the network average power consumption

per area by turning on only a fraction of micro BSs according to an activation ratio. Numerical results demonstrate that the network average power consumption per area can be minimized by optimally tuning the activation ratio.

5.1.2 QoS Constraint

Besides energy efficiency, QoS provisioning is an important issue in HetNets. In [99,103], users' QoS was guaranteed by the consideration that the network rate coverage should not fall below a target value. The ratio between BSs' sum rate and their power consumption in the network was optimized in [97,98,106,107,124–126]. Apart from the rate related QoS metrics, the delay performance attracts more and more attention with the proliferation of real-time multimedia applications. In particular, Zhong *et al.* [127] derived the probability that the mean delay of a packet exceeds a given threshold in a K-tier HetNet. Zhang *et al.* [128] analyzed the impact of the system parameters such as the BS density on the local delay, i.e., the mean number of slots for a packet to be delivered. In [129], the optimal user association and BS resource allocation was obtained to minimize the average packet delay. Nevertheless, [127–129] all assumed consistent interference from other co-channel BSs, due to which a worst case of the delay performance was characterized in [127–129].

This thus motivates us to characterize the delay performance in a more precise manner by capturing the correlation between queuing and interference. In particular, by using the approximation that the BSs in one tier have an identical SIR coverage, the cumulative distribution function (CDF) of the traffic intensity of the BSs in each tier is characterized, based on which the network mean queuing delay is further obtained. We then formulate the QoS constraints by guaranteeing that the network mean queuing delay is lower than a certain value and that the network SIR coverage is higher than a certain value. The threshold value could be determined by QoS requirements. The analytical results show that the constraints of the network mean queuing delay and the network SIR coverage

can be translated into a lower bound fraction of micro BSs that should be switched on.

5.1.3 Universal Frequency Reuse Versus Spectrum Partitioning

Throughout this chapter, universal frequency reuse (UFR) is adopted where BSs of all tiers share the whole spectrum resources. Nevertheless, due to the intensive BS deployment in HetNets, BSs become much closer to each other. The interference in HetNets is thus much more severe than that in traditional cellular networks. As a result, spectrum partitioning (SP) was advocated in a great deal of previous works to mitigate the inter-tier interference in HetNets [68, 118, 121, 122]. In particular, Lin *et al.* [68] maximized the logarithm of users' rate by properly allocating the bandwidth to each tier. Similarly, Ramamonjison *et al.* [118] optimized the bandwidth allocation between macro and micro cells to maximize the area spectral efficiency. To minimize the average downlink user data rate, Bao *et al.* [121, 122] proposed a structured spectrum allocation and user association scheme, and showed that BSs of a tier with higher deployment density should have a higher priority in spectrum allocation. Compared to UFR, it was observed in [68, 118, 121, 122] that the system performance can be significantly improved by a properly spectrum allocation.

As queuing is not assumed in [68, 118, 121, 122], the interference level of a user solely depends on the number of interfering BSs. In consideration of traffic dynamics, nevertheless, the interference is not only determined by the number of BSs but also is closely related to the queuing status of each co-channel BS. This leads to the question that whether SP can still perform better than UFR. To address this issue, this chapter further conducts a comparative study between UFR and SP in terms of the network average power consumption per area and the network SIR coverage. It is found that although the number of interfering BSs increases when reusing the spectrum resources among all tiers, the bandwidth allocated to each tier is enlarged. Therefore, the inter-tier interference will not necessarily deteriorate as it can be mitigated by a better queuing performance of

the BSs. The simulation results show that UFR outperforms SP in terms of both energy efficiency and SIR coverage in the considered scenario.

The contributions of this chapter are summarized as follows.

- With the consideration of dynamical traffic arrivals in a 2-Tier HetNet, the average traffic intensity of each tier is characterized by a set of fixed-point equations, based upon which the existence and uniqueness of the solution is obtained. To numerically solve the fixed-point equations, an iterative method is proposed. By using the approximation that BSs of a tier have the same SIR coverage, the CDF of the traffic intensity of each tier is obtained.

- Based on the characterization of the traffic intensity, a network average power consumption minimization problem under the constraints of the network mean queuing delay and the network SIR coverage is formulated. Numerical results show that if the idle power coefficient is below a certain threshold, the optimal activation ratio should equal the one to minimize the network average power consumption per area. Otherwise, the optimal activation ratio should be obtained according to the QoS constraints.

- On the contrary to previous observations, a comparative study show that UFR outperforms SP in terms of both energy efficiency and SIR coverage in the considered scenario.

The rest of this chapter is organized as follows. The system model is presented in Section 5.2. The BS traffic intensity is characterized in Section 5.3. A network average power consumption optimization problem under the constraints of the network mean queuing delay and network SIR coverage is formulated and studied in Section 5.4. Simulation results are demonstrated in Section 5.5. A comparative study of UFR and SP is studied in Section 5.6. Conclusions are given in Section 5.7.

5.2 System Model

We consider a 2-Tier HetNet where BSs in the k-th tier are spatially distributed as a PPP with the deployment density λ_k^d, $k = 1, 2$. Without loss of generality, Tier-1 BSs are referred to as macro BSs with a higher transmission power P_1 and a lower deployment density λ_1^d, and Tier-2 BSs are referred to as micro BSs with a lower transmission power P_2 and a higher deployment density λ_2^d. To provide global coverage, we assume that all Tier-1 BSs are active while Tier-2 BSs could be switched off with an active ratio ϵ. Therefore, the set of active Tier-1 BSs, Φ_1, has an intensity $\lambda_1 = \lambda_1^d$, and the set of active Tier-2 BSs, Φ_2, has an intensity $\lambda_2 = \epsilon\lambda_2^d$. Mobile users, on the other hand, are modeled by an independent PPP Φ_{u} with an intensity λ_{u}. Each mobile user in the downlink connects to a BS that offers the highest received power. Universal frequency reuse (UFR) is adopted where the whole bandwidth of W is shared by all the BSs across tiers.

Similar to the queuing assumption in Chapter 3 and Chapter 4, it is assumed that the packets randomly arrive and queue in the buffer of its associated BS. BSs then send the packets in the buffer to its associated users in a FIFS fashion. The packet arrival follows a Poisson process with the mean arrival rate γ, and the packet size is exponentially distributed with the mean length L. Note that as UFR is adopted in this chapter, for the typical user associated to a Tier-k BS, all other BSs are potential interferers. To prevent users in poor channel conditions from occupying the spectrum resource, we still assume that for each user, its packet would be dropped if the SIR is below a predetermined threshold τ; otherwise, it is served with a constant rate $W\log_2(1+\tau)$.

5.3 Traffic Intensity Characterization

In this section, we will first study the average traffic intensity of each tier $\bar{\rho}_k$, based on which cumulative distribution function of $\rho_{k,i}$ is obtained.

5.3.1 Average Traffic Intensity

According to the characterization of the average traffic intensity with UFR in Section 2.3.2, we can apply the fixed-point equations (2.26) to our considered 2-Tier HetNet as

$$\bar{\rho}_k = \frac{\gamma \lambda_{\mathrm{u}} L A_k}{\lambda_k W \log_2(1+\tau)} \cdot \frac{1}{A_k \sum_{j=1}^{2} \tilde{\lambda}_j \tilde{P}_j^{-\frac{2}{\alpha}} \bar{\rho}_j Z + 1}, \tag{5.1}$$

where $k \in \{1, 2\}$. It can be seen from (5.1) that the average traffic intensity of each tier, i.e., $\bar{\rho}_1$ and $\bar{\rho}_2$, depends on each other. We will demonstrate how to obtain $\bar{\rho}_k$ in the following.

Corollary 5.1. *The fixed-point equations* (5.1) *have a unique solution within the region* $\bar{\rho}_k > 0$, $k \in \{1, 2\}$.

Proof. According to (5.1), we have

$$\bar{\rho}_2 = \frac{\beta_1}{\theta_1 \bar{\rho}_1} - \frac{A_1 Z \bar{\rho}_1}{\theta_1} - \frac{1}{\theta_1}, \tag{5.2a}$$

and

$$\bar{\rho}_2 = \frac{-(1 + \theta_2 \bar{\rho}_1) + \sqrt{(1 + \theta_2 \bar{\rho}_1)^2 + 4 A_2 Z \beta_2}}{2 A_2 Z}, \tag{5.2b}$$

where

$$\beta_1 = \frac{\gamma \lambda_{\mathrm{u}} L A_1}{\lambda_1 W \log_2(1+\tau)}, \tag{5.3a}$$

$$\theta_1 = A_1 \frac{\lambda_2}{\lambda_1} \cdot \left(\frac{P_2}{P_1}\right)^{\frac{2}{\alpha}} Z, \tag{5.3b}$$

$$\beta_2 = \frac{\gamma \lambda_{\mathrm{u}} L A_2}{\lambda_2 W \log_2(1+\tau)}, \tag{5.3c}$$

$$\theta_2 = A_2 \frac{\lambda_1}{\lambda_2} \cdot \left(\frac{P_1}{P_2}\right)^{\frac{2}{\alpha}} Z. \tag{5.3d}$$

Hence, the solution of (5.1) can be obtained by solving

$$\delta = \frac{\beta_1}{\theta_1\bar{\rho}_1} - \frac{A_1 Z\bar{\rho}_1}{\theta_1} - \frac{1}{\theta_1} - \frac{-\left(1+\theta_2\bar{\rho}_1\right) + \sqrt{\left(1+\theta_2\bar{\rho}_1\right)^2 + 4A_2 Z\beta_2}}{2A_2 Z} = 0. \tag{5.4}$$

Since $\lim_{\bar{\rho}_1\to 0}\delta \to \infty$ and $\lim_{\bar{\rho}_1\to\infty}\delta \to -\infty$, there exists at least one solution to (5.4).

Furthermore, according to (5.2a), we have

$$\frac{d\bar{\rho}_2}{d\bar{\rho}_1} = -\frac{\beta_1}{\theta_1\bar{\rho}_1^2} - \frac{A_1 Z}{\theta_1}, \tag{5.5a}$$

$$\frac{d^2\bar{\rho}_2}{d\bar{\rho}_1^2} = \frac{\beta_1}{2\theta_1\bar{\rho}_1^3} > 0, \tag{5.5b}$$

and according to (5.2b), we have

$$\frac{d\bar{\rho}_2}{d\bar{\rho}_1} = -\frac{\theta_2}{2A_2 Z} - \frac{\theta_2\left(1+\theta_2\bar{\rho}_1\right)}{2A_2 Z\sqrt{\left(1+\theta_2\bar{\rho}_1\right)^2 + 4A_2 Z\beta_2}}, \tag{5.5c}$$

$$\frac{d^2\bar{\rho}_2}{d\bar{\rho}_1^2} = \frac{\theta_2^2}{2A_2 Z}\cdot\frac{4A_2 Z\beta_2}{\left[\left(1+\theta_2\bar{\rho}_1\right)^2 + 4A_2 Z\beta_2\right]^{\frac{3}{2}}} > 0. \tag{5.5d}$$

By combining (5.4) with (5.5), we have

$$\begin{aligned}\frac{d\delta}{d\bar{\rho}_1} &< \lim_{\bar{\rho}_1\to\infty}\left(-\frac{\beta_1}{\theta_1\bar{\rho}_1^2} - \frac{A_1 Z}{\theta_1}\right) - \frac{\theta_2}{2A_2 Z}\lim_{\bar{\rho}_1\to 0}\left(-1 + \frac{1+\theta_2\bar{\rho}_1}{\sqrt{\left(1+\theta_2\bar{\rho}_1\right)^2 + 4A_2 Z\beta_2}}\right)\\ &< -\frac{A_1 Z}{\theta_1} + \frac{\theta_2}{2A_2 Z} = -\frac{\lambda_1}{2\lambda_2}\cdot\left(\frac{P_1}{P_2}\right)^{\frac{2}{\alpha}} < 0,\end{aligned} \tag{5.6}$$

which indicates that (5.4) has a unique solution. □

Generally, (5.1) does not have a closed-form solution, and the average traffic intensity of each tier could be obtained iteratively as

$$\bar{\rho}_1^{(n+1)} = \frac{\gamma\lambda_{\mathrm{u}} L A_1}{\lambda_1 W\log_2\left(1+\tau\right)\left(1 + A_1\bar{\rho}_1^{(n)} Z + A_1\frac{\lambda_2}{\lambda_1}\cdot\left(\frac{P_2}{P_1}\right)^{\frac{2}{\alpha}}\bar{\rho}_2^{(n)} Z\right)} \tag{5.7a}$$

$$\bar{\rho}_2^{(n+1)} = \frac{\gamma \lambda_{\mathrm{u}} L A_2}{\lambda_2 W \log_2(1+\tau)\left(1 + A_2 \bar{\rho}_2^{(n)} Z + A_2 \frac{\lambda_1}{\lambda_2} \cdot \left(\frac{P_1}{P_2}\right)^{\frac{2}{\alpha}} \bar{\rho}_1^{(n)} Z\right)} \tag{5.7b}$$

Note that $\bar{\rho}_1^{(n)}$ and $\bar{\rho}_2^{(n)}$ are the average traffic intensities after n-th iterations. The following lemma proves the convergence of (5.7).

Lemma 4.1. *Solving the fixed-point equations* (5.1) *by the iterative method* (5.7) *converges to a unique set of points, which is the solution of* (5.1).

Proof. By substituting (5.3) into (5.7), and denoting $\delta_1 = A_1 Z$ and $\delta_2 = A_2 Z$, we have

$$\bar{\rho}_1^{(n+1)} = \frac{\beta_1}{1 + \delta_1 \bar{\rho}_1^{(n)} + \theta_1 \bar{\rho}_2^{(n)}}, \tag{5.8a}$$

$$\bar{\rho}_2^{(n+1)} = \frac{\beta_2}{1 + \delta_2 \bar{\rho}_2^{(n)} + \theta_2 \bar{\rho}_1^{(n)}}. \tag{5.8b}$$

Meanwhile, we have shown in Corollary 1 that (5.1) has a unique solution of $\bar{\rho}_1$ and $\bar{\rho}_2$, which is denoted as as $\bar{\rho}_1^*$ and $\bar{\rho}_2^*$, respectively. By initializing the iteration with $\bar{\rho}_1^{(0)} < \bar{\rho}_1^*$ and $\bar{\rho}_2^{(0)} < \bar{\rho}_1^*$, we have

$$\bar{\rho}_1^{(1)} = \frac{\beta_1}{1 + \delta_1 \bar{\rho}_1^{(0)} + \theta_1 \bar{\rho}_2^{(0)}} > \frac{\beta_1}{1 + \delta_1 \bar{\rho}_1^{*} + \theta_1 \bar{\rho}_2^{*}} = \bar{\rho}_1^*, \tag{5.9a}$$

$$\bar{\rho}_2^{(1)} = \frac{\beta_2}{1 + \delta_2 \bar{\rho}_2^{(0)} + \theta_2 \bar{\rho}_1^{(0)}} > \frac{\beta_1}{1 + \delta_2 \bar{\rho}_2^{*} + \theta_2 \bar{\rho}_1^{*}} = \bar{\rho}_2^*, \tag{5.9b}$$

and

$$\bar{\rho}_1^{(2)} = \frac{\beta_1}{1 + \delta_1 \bar{\rho}_1^{(1)} + \theta_1 \bar{\rho}_2^{(1)}} < \frac{\beta_1}{1 + \delta_1 \bar{\rho}_1^{*} + \theta_1 \bar{\rho}_2^{*}} = \bar{\rho}_1^*, \tag{5.9c}$$

$$\bar{\rho}_2^{(2)} = \frac{\beta_2}{1 + \delta_2 \bar{\rho}_2^{(1)} + \theta_2 \bar{\rho}_1^{(1)}} < \frac{\beta_1}{1 + \delta_2 \bar{\rho}_2^{*} + \theta_2 \bar{\rho}_1^{*}} = \bar{\rho}_2^*. \tag{5.9d}$$

Therefore, we conclude without loss of generality that $\bar{\rho}_k^{(2n)} < \bar{\rho}_k^*$ and $\bar{\rho}_k^{(2n+1)} > \rho_k^*$ where $k \in \{1,2\}$. On the other hand, if we set a sufficiently small initial $\bar{\rho}_k^{(0)}$, for example, $\bar{\rho}_1^{(0)} = \bar{\rho}_2^{(0)} = 0$, we have

$$\bar{\rho}_k^{(2)} > 0 = \bar{\rho}_k^{(0)},\ k \in \{1,2\}, \tag{5.10a}$$

and

$$\begin{aligned} \frac{\bar{\rho}_1^{(3)}}{\bar{\rho}_1^{(1)}} &= \frac{1+\delta_1\bar{\rho}_1^{(0)}+\theta_1\bar{\rho}_2^{(0)}}{1+\delta_1\bar{\rho}_1^{(2)}+\theta_1\bar{\rho}_2^{(2)}} < 1, \\ \frac{\bar{\rho}_2^{(3)}}{\bar{\rho}_2^{(1)}} &= \frac{1+\delta_2\bar{\rho}_2^{(0)}+\theta_2\bar{\rho}_1^{(0)}}{1+\delta_2\bar{\rho}_2^{(2)}+\theta_2\bar{\rho}_1^{(2)}} < 1. \end{aligned} \tag{5.10b}$$

By assuming $\bar{\rho}_k^{(2n+2)}/\bar{\rho}_k^{(2n)} > 1$ and $\bar{\rho}_k^{(2n+1)}/\bar{\rho}_k^{(2n-1)} < 1$ hold for any given n-th iteration, we further have

$$\begin{aligned} \frac{\bar{\rho}_1^{(2n+3)}}{\bar{\rho}_1^{(2n+1)}} &= \frac{1+\delta_1\bar{\rho}_1^{(2n)}+\theta_1\bar{\rho}_2^{(2n)}}{1+\delta_1\bar{\rho}_1^{(2n+2)}+\theta_1\bar{\rho}_2^{(2n+2)}} < 1, \\ \frac{\bar{\rho}_2^{(2n+3)}}{\bar{\rho}_2^{(2n+1)}} &= \frac{1+\delta_2\bar{\rho}_2^{(2n)}+\theta_2\bar{\rho}_1^{(2n)}}{1+\delta_2\bar{\rho}_2^{(2n+1)}+\theta_2\bar{\rho}_1^{(2n+1)}} < 1, \end{aligned} \tag{5.11a}$$

and

$$\begin{aligned} \frac{\bar{\rho}_1^{(2n+4)}}{\bar{\rho}_1^{(2n+2)}} &= \frac{1+\delta_1\bar{\rho}_1^{(2n+1)}+\theta_1\bar{\rho}_2^{(2n+1)}}{1+\delta_1\bar{\rho}_1^{(2n+3)}+\theta_1\bar{\rho}_2^{(2n+3)}} > 1, \\ \frac{\bar{\rho}_2^{(2n+4)}}{\bar{\rho}_2^{(2n+2)}} &= \frac{1+\delta_2\bar{\rho}_2^{(2n+1)}+\theta_2\bar{\rho}_1^{(2n+1)}}{1+\delta_2\bar{\rho}_2^{(2n+3)}+\theta_2\bar{\rho}_1^{(2n+3)}} > 1, \end{aligned} \tag{5.11b}$$

which also hold for the $(n+1)$-th iteration. Therefore, by combining $\bar{\rho}_k^{(2n)} < \bar{\rho}_k^*$ and $\bar{\rho}_k^{(2n+1)} > \bar{\rho}_k^*$, it can be concluded that $\bar{\rho}_k^{(2n)}$ and $\bar{\rho}_k^{(2n+1)}$ would converge as $n \to \infty$.

Finally, if $\bar{\rho}_k^{(2n)}$ and $\bar{\rho}_k^{(2n+1)}$ converge to different points, i.e., $\lim_{n\to\infty} \bar{\rho}_k^{(2n)} = \bar{\rho}_k^{1,*}$ and $\lim_{n\to\infty} \bar{\rho}_k^{(2n+1)} = \bar{\rho}_k^{2,*}$ where $\bar{\rho}_k^{1,*} \neq \bar{\rho}_k^{2,*}$, (5.1) would hold for both $\rho_k^{1,*}$ and $\rho_k^{2,*}$, which is contradictory to Corollary 5.1. Hence, $\bar{\rho}_k^{1,*} = \bar{\rho}_k^{2,*} = \bar{\rho}_k^*$, indicating that $\bar{\rho}_k^{(2n)}$ and $\bar{\rho}_k^{(2n+1)}$ would converge to $\bar{\rho}_k^*$.

□

In a lightly-loaded network scenario where the mean packet arrival of each user γ is low, (5.1) can be written as

$$\bar{\rho}_k \approx \frac{\gamma \lambda_{\mathrm{u}} L A_k}{\lambda_k W \log_2(1+\tau)} \cdot \left(1 - A_k \sum_{j=1}^{2} \lambda_j P_j^{-\frac{2}{\alpha}} \bar{\rho}_j Z\right), \tag{5.12}$$

by using the approximation $1/(1+x) \approx 1-x$ when x is small. According to (5.12), we then have the explicit expressions of the average traffic intensity of each tier, i.e.,

$$\bar{\rho}_1 = \frac{\theta_1 - \left(A_2 Z + \frac{1}{\beta_2}\right)}{\theta_1\theta_2 - \left(A_1 Z + \frac{1}{\beta_1}\right)\left(A_2 Z + \frac{1}{\beta_2}\right)}, \tag{5.13a}$$

$$\bar{\rho}_2 = \frac{\theta_2 - \left(A_1 Z + \frac{1}{\beta_1}\right)}{\theta_1\theta_2 - \left(A_1 Z + \frac{1}{\beta_1}\right)\left(A_2 Z + \frac{1}{\beta_2}\right)}, \tag{5.13b}$$

where β_1, β_2, θ_1 and θ_2 are given in (5.3).

To this end, we have obtained the average traffic intensity $\bar{\rho}_k$ of each tier. We will further characterize the cumulative distribution function of $\rho_{k,i}$ in the following subsection.

5.3.2 Cumulative Distribution Function of Traffic Intensity

According to (2.5), the traffic intensity $\rho_{k,i}$ relies on two random variables, i.e., the number of associated users $N_{k,i}$ and the SIR coverage $S_{k,i}$. As $N_{k,i}$ and $S_{k,i}$ depends on each other, it is difficult to characterize the distribution of $\rho_{k,i}$ exactly. To simplify analysis, we use the approximation that $S_{k,i}$ is identical across BSs of one tier and equals the SIR coverage of this tier, i.e., $S_{k,i} = S_k$. The traffic intensity can then be written as

$$\rho_{k,i} = \frac{\gamma N_{k,i} S_k L}{W \log_2(1+\tau)}. \tag{5.14}$$

It can be clearly seen from (5.14) that the distribution of the traffic intensity $\rho_{k,i}$ now is solely determined by the number of associated users $N_{k,i}$, which is a discrete random

variable. The following corollary gives the probability mass function of the traffic intensity $\rho_{k,i}$.

Corollary 5.2. *The probability mass function of the traffic intensity $\rho_{k,i}$ is given by*

$$\Pr\left[\rho_{k,i}=\frac{\gamma S_k L n}{W\log_2\left(1+\tau\right)}\right]=3.5^{3.5}\frac{\lambda_{\mathrm{u}}^n}{n!}\left(\frac{\lambda_k}{A_k}\right)^{3.5}\frac{\prod_{i=0}^{n-1}3.5+i}{\left(\lambda_{\mathrm{u}}+3.5\frac{\lambda_k}{A_k}\right)^{n+3.5}}. \tag{5.15}$$

Proof. Recall that users form a PPP distribution with the intensity λ_{u}. With a given association area T_k, the distribution of the number of associated users $N_{k,i}$ is given by

$$\Pr\left[N_{k,i}=n\left|T_k\right.\right]=\frac{\left(\lambda_{\mathrm{u}}T_k\right)^n e^{-\lambda_{\mathrm{u}}T_k}}{n!}. \tag{5.16}$$

Meanwhile, according to [130], the association area of a random Tier-k BS, T_k, follows the following probability density function (PDF)

$$f_{T_k}\left(T_k\right)=\frac{3.5^{3.5}\lambda_k}{\Gamma\left(3.5\right)A_k}\left(\frac{\lambda_k T_k}{A_k}\right)^{2.5}e^{-3.5\frac{\lambda_k T_k}{A_k}}, \tag{5.17}$$

where $\Gamma\left(x\right)=\int_0^\infty t^{x-1}e^{-t}dt$. Therefore, the probability mass function (PMF) of $N_{k,i}$ can be written as

$$\begin{aligned}\Pr\left[N_{k,i}=n\right]&=\int_0^\infty\Pr\left[N_{k,i}=n\left|T_k\right.\right]\cdot f_{T_k}\left(T_k\right)dT_k\\&=\int_0^\infty\frac{\left(\lambda_{\mathrm{u}}T_k\right)^n e^{-\lambda_{\mathrm{u}}T_k}}{n!}\cdot\frac{3.5^{3.5}\lambda_k}{\Gamma\left(3.5\right)A_k}\left(\frac{\lambda_k T_k}{A_k}\right)^{2.5}e^{-3.5\frac{\lambda_k T_k}{A_k}}dT_k\\&=\frac{3.5^{3.5}\lambda_{\mathrm{u}}^n}{\Gamma\left(3.5\right)n!}\left(\frac{\lambda_k}{A_k}\right)^{3.5}\cdot\frac{\prod_{i=o}^{n-1}3.5+i}{\left(\lambda_{\mathrm{u}}+3.5\frac{\lambda_k}{A_k}\right)^n}\int_0^\infty T_k^{2.5}e^{-\left(\lambda_{\mathrm{u}}+3.5\frac{\lambda_k}{A_k}\right)T_k}dT_k\\&=3.5^{3.5}\frac{\lambda_{\mathrm{u}}^n}{n!}\left(\frac{\lambda_k}{A_k}\right)^{3.5}\frac{\prod_{i=0}^{n-1}3.5+i}{\left(\lambda_{\mathrm{u}}+3.5\frac{\lambda_k}{A_k}\right)^{n+3.5}}.\end{aligned} \tag{5.18}$$

Finally, (5.15) can be obtained by combining (5.14) with (5.18). □

According to Corollary 5.2, the cumulative distribution function (CDF) of $\rho_{k,i}$ can then be written as

$$\begin{aligned}\Pr\left[\rho_{k,i} < \rho\right] &= \Pr\left[\frac{\gamma S_k L N_{k,i}}{W \log_2\left(1+\tau\right)} < \rho\right] \\ &= \sum_{n=0}^{N} 3.5^{3.5} \frac{\lambda_{\mathrm{u}}^n}{n!} \left(\frac{\lambda_k}{A_k}\right)^{3.5} \frac{\prod_{i=0}^{n-1} 3.5+i}{\left(\lambda_{\mathrm{u}} + 3.5\frac{\lambda_k}{A_k}\right)^{n+3.5}},\end{aligned} \tag{5.19}$$

where $N = \left\lfloor \frac{W\log_2(1+\tau)}{\gamma S_k L} \right\rfloor$ is the maximum number of associated users as we assume that each BS is unsaturated. It will be demonstrated in Section 5.5 that the analysis in (5.19) is close to the simulation results, indicating that the approximation achieves a good accuracy.

5.4 QoS Constrained Network Average Power Consumption Optimization

It has been shown in Section 5.3 that the traffic intensity depends on the density of active micro BSs, i.e., $\lambda_2 = \epsilon\lambda_2^d$. As the power consumption of a BS varies significantly between the busy and the idle states, the network energy efficiency is critically determined by the activation ratio ϵ of micro BSs. Therefore, we formulate a network average power consumption minimization problem under the QoS constraints of the network SIR coverage and the network mean queuing delay as

$$\min_{\epsilon} \; P, \tag{5.20a}$$

$$\text{s.t.} \;\; \mathrm{S} > \hat{\mathrm{S}}, \tag{5.20b}$$

$$D < \hat{D}. \tag{5.20c}$$

where P denotes the network average power consumption per area; Constraint (5.20b) indicates that the network SIR coverage S does not drop below a certain threshold $\hat{S}$; Constraint (5.20c) guarantees that the network mean queuing delay D does not exceed a target value $\hat{D}$. We will further derive the network average power consumption per area P, the network SIR coverage S, and the network mean queuing delay D in the following.

5.4.1 Performance Metrics

Network Average Power Consumption Per Area

According to [95], the power consumption of an active Tier-k BS in busy state $P_{k,\mathrm{T}}$ can be written as $P_{k,\mathrm{T}} = P_{k,s} + \Delta_k P_k$ where $P_{k,s}$ denotes the power consumption of its signal processing and battery leakage, and Δ_k denotes its cooling and feeder loss. The BS in the idle state would consume less energy than it does in the busy state. Therefore, we use an idle state coefficient $\eta_k < 1$ to denote the power consumption ratio of an active Tier-k BS between the idle state and the busy state. The power consumption of an active Tier-k BS in the idle state can then be written as $P_{k,\mathrm{I}} = \eta_k P_{k,\mathrm{T}}$. In consideration that if the power consumption of one active BS is higher in the busy state, its power consumption in the idle state will also be higher, the coefficient η_k is assumed to be identical across tiers, i.e., $\eta_1 = \eta_2 = \eta$, for simplicity.

Therefore, the average power consumption of an active Tier-k BS can be written as

$$\begin{aligned} P_{k,\mathrm{av}} &= E\left[\rho_{k,i} P_{k,\mathrm{T}} + P_{k,\mathrm{I}}\left(1-\rho_{k,i}\right)\right] = P_{k,\mathrm{T}} E\left[\rho_{k,i}\right] + \eta P_{k,\mathrm{T}} - \eta P_{k,\mathrm{T}} E\left[\rho_{k,i}\right] \\ &= \left(1-\eta\right) P_{k,\mathrm{T}} \bar{\rho}_k + \eta P_{k,\mathrm{T}}. \end{aligned} \tag{5.21}$$

The network average power consumption per area P can then be written as

$$\begin{aligned} P &= \sum_{k=1}^{2} \lambda_k \left(1-\eta\right) P_{k,\mathrm{T}} \bar{\rho}_k + \lambda_k \eta P_{k,\mathrm{T}} \\ &= \lambda_1^d P_{1,\mathrm{T}} \left[\left(1-\eta\right)\bar{\rho}_1 + \eta\right] + \epsilon \lambda_2^d P_{2,\mathrm{T}} \left[\left(1-\eta\right)\bar{\rho}_2 + \eta\right], \end{aligned} \tag{5.22}$$

By having a larger activation ratio, the average busy probability of each active BS can be reduced by offloading the traffic pressure to more active Tier-2 BSs, leading to a lower average power consumption of an individual BS. On the other hand, the network power consumption increases as the number of active Tier-2 BSs increases. Therefore, there exists an optimal activation ratio $\bar{\epsilon}$ to minimize the network average power consumption per area.

QoS Constraints

In this subsection, we first characterize the constraint of the network SIR coverage. According to (2.22), the network SIR coverage can be obtained as

$$S = \sum_{k=1}^{2} A_k S_k = \sum_{k=1}^{2} \frac{A_k}{1 + A_k \sum\limits_{j=1}^{2} \tilde{\lambda}_j \tilde{P}_j^{2/\alpha} \bar{\rho}_j Z}. \tag{5.23}$$

(5.20b) can then be written as

$$\sum_{k=1}^{2} \frac{A_k}{1 + A_k \sum\limits_{j=1}^{2} \tilde{\lambda}_j \tilde{P}_j^{2/\alpha} \bar{\rho}_j Z} > \hat{S}. \tag{5.24}$$

Numerical results show that (5.24) can be converted to a lower bound of Tier-2 BS activation ratio $\epsilon^s_{\min}$. Fig. 5.1 demonstrates how the lower bound of the activation ratio $\epsilon^s_{\min}$ varies with the network SIR coverage threshold $\hat{S}$ with various values of the mean bit arrival rate per area $\gamma\lambda_{\mathrm{u}}L$. It can be observed from Fig. 5.1 that with a given $\gamma\lambda_{\mathrm{u}}L$, $\epsilon^s_{\min}$ monotonically increases as the threshold $\hat{S}$ increases. Intuitively, as we consider queuing in the BSs, the network SIR coverage (5.23) is not only determined by the number of interferers but also is affected by the queuing performance of each BS. Although the number of interfering sources increases by activating more Tier-2 BSs, the probability that one BS is in the busy state can be significantly reduced, which offsets the effect of the increment of the number of BSs. Therefore, as the threshold $\hat{S}$ increases, the lower

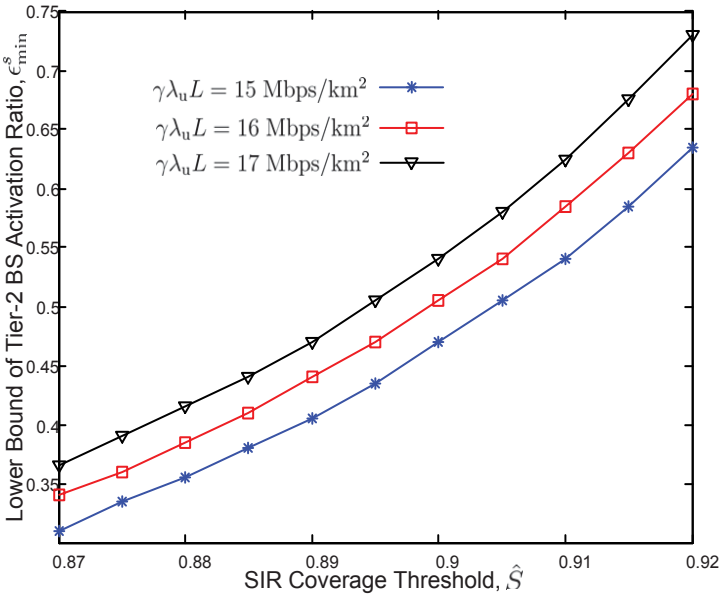


Figure 5.1: The lower bound of Tier-2 BS activation ratio $\epsilon^s_{\min}$ versus the network SIR coverage threshold $\hat{S}$. The system parameters can be found in Table I.

bound of the activation ratio $\epsilon^s_{\min}$ becomes larger, which indicates that more Tier-2 BSs should be switched on.

In the following, we further study the constraint of the network mean queuing delay. Since each BS is modeled as a M/D/1 queuing system, the mean queuing delay of Tier k can be obtained as

$$D_k = E\left[\frac{1}{\mu_k\left(1-\rho_{k,i}\right)}\right]. \tag{5.25}$$

According to Corollary 5.2, (5.25) can be written as

$$\begin{aligned} D_k &= \sum_n \frac{L}{W\log_2\left(1+\tau\right)-\gamma S_k L n}\Pr\left[\rho_{k,i}=\frac{\gamma S_k L n}{W\log_2\left(1+\tau\right)}\right] \\ &= \sum_{n=0}^{N} \frac{3.5^{3.5}L}{W\log_2\left(1+\tau\right)-\gamma S_k L n}\cdot\frac{\lambda_u^n}{n!}\left(\frac{\lambda_k}{A_k}\right)^{3.5}\frac{\prod_{i=0}^{n-1}3.5+i}{\left(\lambda_u+3.5\frac{\lambda_k}{A_k}\right)^{n+3.5}}. \end{aligned} \tag{5.26}$$

The network mean queuing delay can thus be written as

$$D = \sum_{k=1}^{2}\frac{\lambda_k}{\lambda_1+\lambda_2}D_k = \frac{\lambda_1^d}{\lambda_1^d+\epsilon\lambda_2^d}D_1 + \frac{\epsilon\lambda_2^d}{\lambda_1^d+\epsilon\lambda_2^d}D_2, \tag{5.27}$$

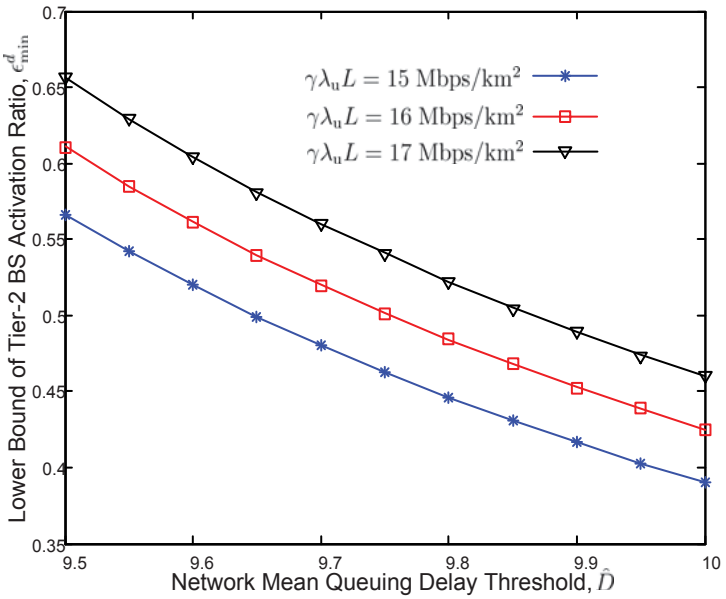


Figure 5.2: The lower bound of Tier-2 BS activation ratio subject to the network mean queuing delay threshold $\hat{D}$. The system parameters can be found in Table I.

and the constraint (5.20c) is then given by

$$\frac{\lambda_1^d}{\lambda_1^d + \epsilon\lambda_2^d} D_1 + \frac{\epsilon\lambda_2^d}{\lambda_1^d + \epsilon\lambda_2^d} D_2 < \hat{D}. \tag{5.28}$$

Similarly, numerical results demonstrate that the constraint (5.28) can be translated into another lower bound of the activation ratio $\epsilon^d_{\min}$. Fig. 5.2 demonstrates how the lower bound $\epsilon^d_{\min}$ varies with the network mean queuing delay threshold $\hat{D}$ with various values of the mean bit arrival rate per area $\gamma\lambda_u L$. It can be observed from Fig. 5.2 that the lower bound $\epsilon^d_{\min}$ decreases as the threshold $\hat{D}$ increases. Intuitively, as more Tier-2 BSs are activated, the probability that a BS is in the idle state decreases. The queuing condition of the network can thus be improved with a larger $\epsilon^d_{\min}$, indicating a lower network mean queuing delay. Therefore, as the network mean queuing delay threshold $\hat{D}$ decreases, more Tier-2 BSs needs to be activated.

5.4.2 QoS Constrained Network Average Power Consumption Minimization

By combining (5.22), (5.24) and (5.28), the QoS constrained network average power consumption optimization problem can be rewritten as

$$\min_{\epsilon} \lambda_1^d P_{1,\mathrm{T}} \left[(1-\eta)\,\bar{\rho}_1 + \eta\right] + \epsilon\lambda_2^d P_{2,\mathrm{T}} \left[(1-\eta)\,\bar{\rho}_2 + \eta\right], \tag{5.29a}$$

$$\text{s.t.} \quad \sum_{\mathrm{k}=1}^{2} \frac{\mathrm{A_k}}{1 + \mathrm{A_k} \sum\limits_{\mathrm{j}=1}^{2} \tilde{\lambda}_{\mathrm{j}} \tilde{\mathrm{P}}_{\mathrm{j}}^{2/\alpha} \bar{\rho}_{\mathrm{j}} \mathrm{Z}} > \hat{\mathrm{S}}, \tag{5.29b}$$

$$\frac{\lambda_1^d}{\lambda_1^d + \epsilon\lambda_2^d} D_1 + \frac{\epsilon\lambda_2^d}{\lambda_1^d + \epsilon\lambda_2^d} D_2 < \hat{D}. \tag{5.29c}$$

Since $\bar{\rho}_k$ does not have an explicit expression, it is difficult to obtain a closed-form solution of the optimization problem (5.29). The optimal Tier-2 BS activation ratio can be obtained numerically. Recall that the constraints (5.29b) and (5.29c) can be translated into $\epsilon > \epsilon_{\min}^s$ and $\epsilon > \epsilon_{\min}^d$. The QoS constraints in (5.29) is thus equivalent to $\epsilon > \epsilon_{\min} = \max\{\epsilon_{\min}^{\mathrm{d}}, \epsilon_{\min}^{\mathrm{s}}\}$. On the other hand, it has been shown in Section 5.4.1 that there exists an optimal activation ratio $\bar{\epsilon}$ to minimize the network average power consumption per area P. Therefore, if $\bar{\epsilon} > \epsilon_{\min}$, the optimal solution of (5.29) can be obtained as $\epsilon^* = \bar{\epsilon}$. Otherwise, as the network average power consumption per area P monotonically increases when $\epsilon > \epsilon_{\min}$, the optimal solution is given by $\epsilon^* = \epsilon_{\min}$. By combining the above two cases, the optimal solution of (5.29) can be written as $\epsilon^* = \max\{\bar{\epsilon}, \epsilon_{\min}\} = \max\{\bar{\epsilon}, \epsilon_{\min}^{\mathrm{d}}, \epsilon_{\min}^{\mathrm{s}}\}$.

5.5 Simulation Results

In this section, we will demonstrate the simulation results to verify the proceeding analysis in Section 5.3 and Section 5.4. The locations of BSs and users are distributed as

Table 5.1: Simulation Parameters

Parameter	Value
User Density λ_{u}	10^{-3} m^{-2}
Tier-1 BS Deployment Density λ_1^d	$1.5 * 10^{-5}$ m^{-2}
Tier-2 BS Deployment Density λ_2^d	$1.5 * 10^{-4}$ m^{-2}
Tier-1 BS Transmission Power P_1	20 W
Tier-2 BS Transmission Power P_2	6 W
Tier-1 BS Fixed Power Consumption $P_{1,s}$	100 W
Tier-2 BS Fixed Power Consumption $P_{2,s}$	25 W
Tier-1 Power Consumption Coefficient Δ_1	4.7
Tier-2 Power Consumption Coefficient Δ_2	2.6
Total Bandwidth W	12 MHz
Path-Loss Coefficient α	4
SIR Threshold τ	1
Mean Packet Length L	0.01 Mb

independent PPPs over a square region of size 5×5 km^2. Monte Carlo simulations are performed over different topologies. The system parameters are shown in Table 5.1.

5.5.1 Traffic Intensity

It has been shown in Section 5.3 that the average traffic intensity of each tier is uniquely determined by a set of fixed-point equations, and can be obtained numerically by an iterative approach. Fig. 5.3 demonstrates how the average traffic intensity $\bar{\rho}_k$ varies with the Tier-2 BS activation ratio ϵ. It can be observed from Fig. 5.3 that the average traffic intensity of each tier $\bar{\rho}_k$ decreases as the activation ratio ϵ increases or as the mean packet arrival rate γ of each user decreases. Intuitively, with a higher ϵ or a lower γ, the queuing performance of the BSs of both tiers improves due to a lower aggregate packet

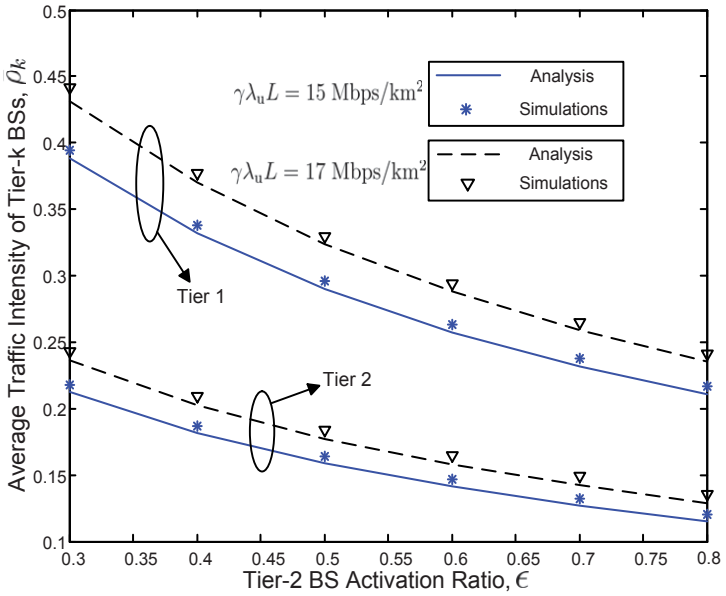


Figure 5.3: Average traffic intensity $\bar{\rho}_k$ versus the Tier-2 BS activation ratio ϵ.

arrival rate of each individual BS. Therefore, BSs have a higher probability of being idle. Simulation results match with the analysis well, which confirms that it serves as a good approximation to regard the locations of busy BSs as a homogeneous thinned PPP.

Recall that in Section 5.3 the CDF of the traffic intensity $\rho_{k,i}$ of each BS has been derived as (5.3.2). Fig. 5.4 illustrates the CDF of the traffic intensity $\Pr[\rho_{k,i} < \rho]$. It can be observed from Fig. 5.4 the simulation results are close to the analysis. However, due to the approximation of replacing the SIR coverage of each individual BS with the SIR coverage of its tier, there exists a small gap between the simulation and analytical results.

5.5.2 Performance Metrics

The network SIR coverage has been derived as (5.23) in Section 5.4.1. Fig. 5.5(a) shows how the network SIR coverage S varies with the Tier-2 BS activation ratio ϵ with various values of the mean bit arrival rate per area $\gamma\lambda_u L$. It can be observed that the network

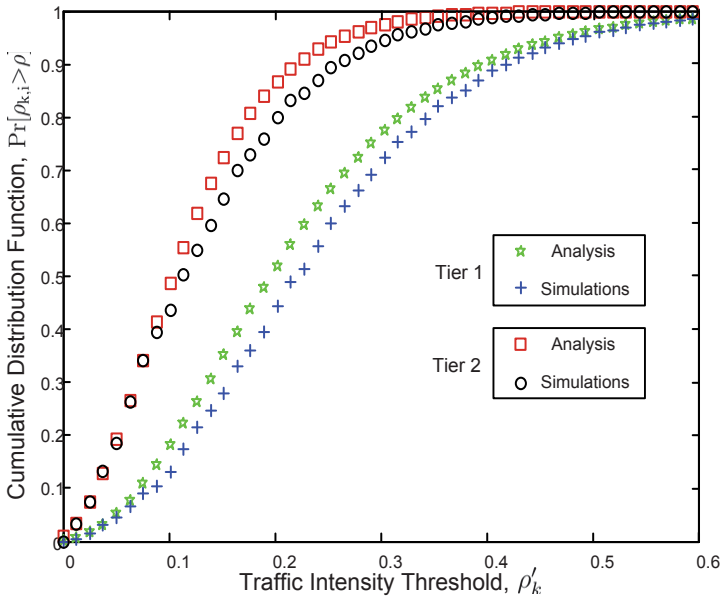
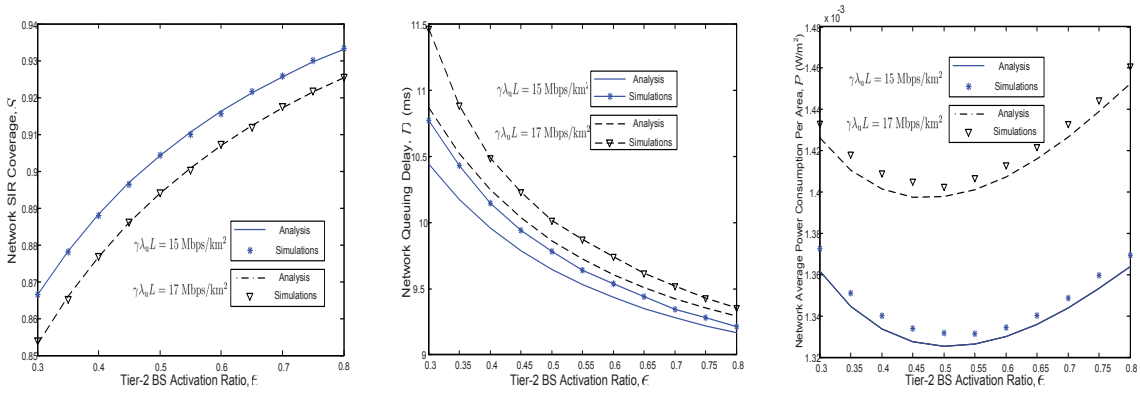


Figure 5.4: Cumulative distribution function of the traffic intensity $\Pr[\rho_{k,i}<\rho]$. $\gamma\lambda_{\mathrm{u}}L$=17 Mbps/km^2.

SIR coverage increases with the activation ratio ϵ. Intuitively, although the number of potential interferers becomes higher with a larger Tier-2 BS activation ratio ϵ, the traffic load can be effectively balanced by the additional activated Tier-2 BSs, in which case BSs of all tiers have a higher probability of being idle and thus are less likely to interfere with each other. Hence, the network SIR coverage increases with the activation ratio ϵ. Furthermore, Fig. 5.5(a) demonstrates that with a given activation ratio ϵ, the network SIR coverage decreases as the mean bit arrival rate per area $\gamma\lambda_{\mathrm{u}}L$ increases. With a higher mean bit arrival rate per area, BSs are more likely to be busy transmitting packets to their users, which leads to a worse network SIR coverage.

On the other hand, the network mean queuing delay has been derived as (5.27) in Section 5.4.1. Fig. 5.5(b) demonstrates how the network mean queuing delay D varies with the Tier-2 BS activation ratio ϵ with various values of the mean bit arrival rate per area $\gamma\lambda_{\mathrm{u}}L$. It can be observed from Fig. 5.5(b) that the network mean queueing delay D decreases as the activation ratio ϵ increases or as the mean bit arrival rate per area $\gamma\lambda_{\mathrm{u}}L$



(a) (b) (c)

Figure 5.5: QoS constrained power consumption optimization problem. $\eta = 0.08$. (a) Network SIR coverage S versus the Tier-2 BS activation ratio ϵ. (b) Network mean queuing delay D versus the Tier-2 BS activation ratio ϵ. (c) Network average power consumption per area P versus the Tier-2 BS activation ratio ϵ.

decreases due to a better queuing performance of each BS. A small gap can be observed from Fig. 5.5(b), which diminishes as ϵ increases. For instance, if $\gamma\lambda_{\mathrm{u}}L = 17\mathrm{Mbps/km}^2$, the gap becomes less than 0.5ms when ϵ exceeds 0.4.

Based on the average traffic intensity $\bar{\rho}_k$, the network average power consumption per area P has been derived as (5.22) in Section 5.4.1. Fig. 5.5(c) illustrates how P varies with the Tier-2 BS activation ratio ϵ. It can be observed from the Fig. 5.5(c) that when the activation ratio ϵ is small, the average network power consumption per area P decreases as ϵ increases; when ϵ becomes large, nevertheless, the average network power consumption per area P increases as ϵ increases. Intuitively, when ϵ is small, the load pressure of the network can be effectively balanced by switching on more Tier-2 BSs. BSs of both tiers thus are more likely to be idle, which improves the network energy efficiency. However, as the ϵ further increases, too many Tier-2 BSs are activated to consume energy. It can be observed from Fig. 5.5(c) that by carefully choosing the Tier-2 BS activation ratio, the average network power consumption per area can be optimized. For example, when the mean bit arrival rate per area $\gamma\lambda_{\mathrm{u}}L = 15\mathrm{Mbps/km}^2$, the optimal activation ratio is given

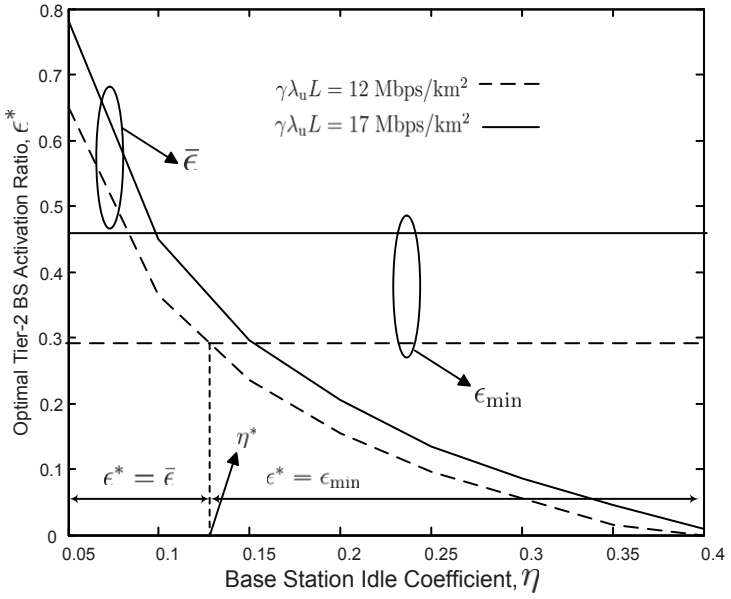


Figure 5.6: Optimal Tier-2 BS activation ratio ϵ^* versus the idle state coefficient η. $\hat{S} = 0.85$ and $\hat{D} = 10$ ms.

by $\bar{\epsilon} = 0.5$.

5.5.3 Optimal Activation Ratio

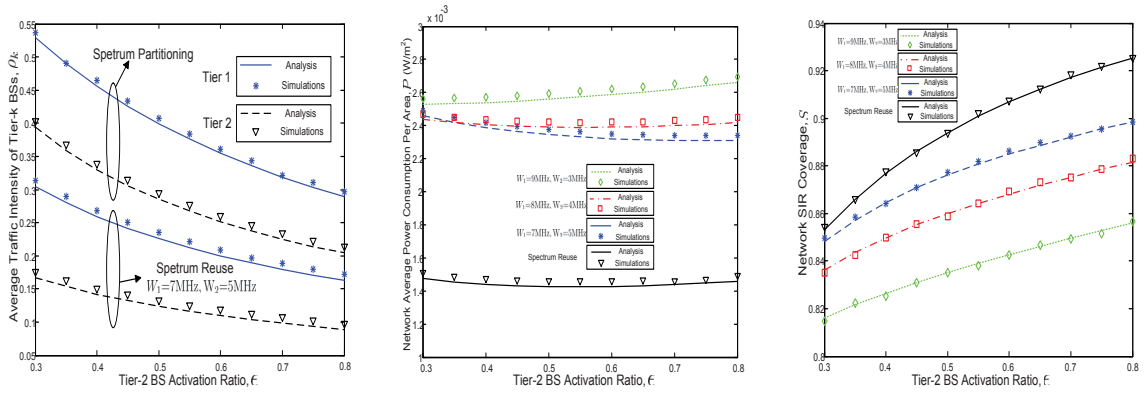
Recall in Section 5.4.1 that the optimal activation ratio of the QoS constrained power consumption optimization problem (5.29) can be written as $\epsilon^* = \max\{\epsilon_{\min}, \bar{\epsilon}\}$ where $\epsilon_{\min}$ denotes the lower bound of the activation ratio subject to the constraints (5.29b) and (5.29c), and $\bar{\epsilon}$ denotes the optimal activation ratio to minimize the network average power consumption per area P. As the BS power consumption in the idle state is critically determined by the idle state coefficient η, Fig. 5.6 demonstrates how the optimal activation ratio ϵ^* varies with the coefficient η with various values of the mean bit arrival rate per area $\gamma\lambda_u L$. It can be observed from Fig. 5.6 that both $\bar{\epsilon}$ and $\epsilon_{\min}$ increase as $\gamma\lambda_u L$ increases. Intuitively, with a higher mean bit arrival rate per area $\gamma\lambda_u L$, the active BSs would be more likely to be busy, leading to a worse QoS and a lower energy efficiency. Therefore, larger $\bar{\epsilon}$ and $\epsilon_{\min}$ are preferred to improve the queuing performance and the

energy efficiency, respectively. As a result, the optimal solution $\epsilon^*= \max\{\bar{\epsilon}, \epsilon_{\min}\}$ increases as γ increases.

On the other hand, it can be observed that $\bar{\epsilon}$ decreases as η increases. As $\epsilon_{\min}$ does not depend on the idle state coefficient η, there exists a certain threshold η^*, such that if $\eta > \eta^*$, we have $\bar{\epsilon} > \epsilon_{\min}$ and $\epsilon^* = \bar{\epsilon}$; otherwise, we have $\bar{\epsilon} < \epsilon_{\min}$ and $\epsilon^* = \epsilon_{\min}$. Intuitively, when η is small, the BS power consumption in the idle state is low. The average power consumption of each individual BS could be effectively reduced by switching on more Tier-2 BSs. Therefore, the optimal activation ratio $\bar{\epsilon}$ to minimize the network average power consumption exceeds its lower bound $\epsilon_{\min}$ to guarantee users' QoS. When η is large, nevertheless, the BS power consumption in the idle state should sharply increase by activating too many Tier-2 BSs. $\bar{\epsilon}$ thus tends to be small to save energy. In this case, $\bar{\epsilon}$ could be lower than $\epsilon_{\min}$, and the optimal fraction of Tier-2 BS to switch on should be $\epsilon^* = \epsilon_{\min}$.

5.6 Comparison of Universal Frequency Reuse and Spectrum Partitioning

Throughout this chapter, universal frequency reuse (UFR) is assumed where BSs of both tiers could interfere with each other. A great deal of previous studies, nevertheless, proposed spectrum partitioning (SP) to mitigate the inter-tier interference in HetNets. As it was assumed in these studies that one BS is transmitting all the time, the interference is only determined by the density of interfering BSs. By considering queuing, the interference level depends on both the number of co-channel BSs and the queuing status of these BSs. As a result, the interference will not necessarily deteriorate by adopting UFR due to a better queuing performance of the BSs. Therefore, we conduct a comparative study between UFR and SP in this section to see whether SP can still perform better in terms



(a) (b) (c)

Figure 5.7: Comparison between universal frequency reuse and spectrum partitioning. $\eta = 0.08$ and $\gamma\lambda_{\mathrm{u}}L = 17\mathrm{Mbps/km}^2$ (a) Average traffic intensity $\bar{\rho}_k$ versus the Tier-2 BS activation ratio ϵ. (b) Network average power consumption per area P versus the Tier-2 BS activation ratio ϵ. (c) Network SIR coverage S versus the Tier-2 BS activation ratio ϵ.

of the network energy efficiency and SIR coverage. Note that with SP in a 2-tier HetNet, the whole bandwidth W is divided into two orthogonal bands W_1 and W_2. Detailed derivations of the network average power consumption per area and the network SIR coverage under the assumption of SP can be found in Chapter 4 and are omitted here.

Fig. 5.7(a) compares the average traffic intensity $\bar{\rho}_k$ with UFR and SP, respectively. It can be observed from Fig. 5.7(a) that the average traffic intensities of both tiers, i.e., $\bar{\rho}_1$ and $\bar{\rho}_2$, can be significantly reduced by adopting UFR. Intuitively, by sharing the whole spectrum across both tiers, the service rate of each BS is higher, leading to a lower busy probability of each individual BS. As a result, it can be observed from Fig. 5.7(b) that the network average power consumption per area P can be greatly reduced by adopting UFR, which indicates that UFR can achieve better energy efficiency over SP.

Fig. 5.7(c) further illustrates the network SIR coverage S under the assumption of UFR and SP, respectively. It can be observed from Fig. 5.7(c) that by adopting UFR, the network SIR coverage S even performs better than that by adopting SP. Intuitively,

although the inter-tier interference is incurred by reusing the spectrum resources, the average BS busy probability $\bar{\rho}_k$ is reduced simultaneously due to a higher bandwidth. Therefore, the inter-tier interference can be mitigated by the decrement of $\bar{\rho}_k$, due to which the SIR coverage can be improved. It can be concluded from Fig. 5.7(b) and Fig. 5.7(c) that in the considered scenario the strategy of UFR is preferred over SP.

Note that power control can be applied to SP to reduce the power consumption. In particular, the transmission power of each tier can be optimally tuned such that the network average power consumption per area can be minimized. For the sake of comparison between UFR and SP, we assume that the average transmission power per area of SP equals that of UFR, which is given by $\lambda_1 P_1 + \epsilon\lambda_2 P_2 = (300 + 900\epsilon)\,\mathrm{W/km^2}$ according to Table 5.1. For SP with power control, the optimal transmission power of each tier can then be obtained by solving the following optimization problem

$$\min_{\{P_1,P_2\}} P, \tag{5.30a}$$

$$\text{s.t. } \lambda_1 P_1 + \epsilon\lambda_2 P_2 = (300 + 900\epsilon)\,\mathrm{W/km^2}, \tag{5.30b}$$

where P is given by (4.5). Fig. 5.8 illustrates how the network minimum average power consumption per area for SP with power control, i.e., (5.30a), varies with the activation ratio ϵ. By comparing Fig. 5.8 with Fig. 5.7, it can be observed that although the network average power consumption per area P with SP is reduced by applying power control, it is still higher than that of UFR. Therefore, it can be concluded that UFR does have better performance over SP even if power control is applied by considering queuing.

5.7 Conclusions

This chapter has studied how to improve energy efficiency by optimally activating a fraction of micro BSs in a 2-Tier HetNet under the QoS constraints of the network SIR

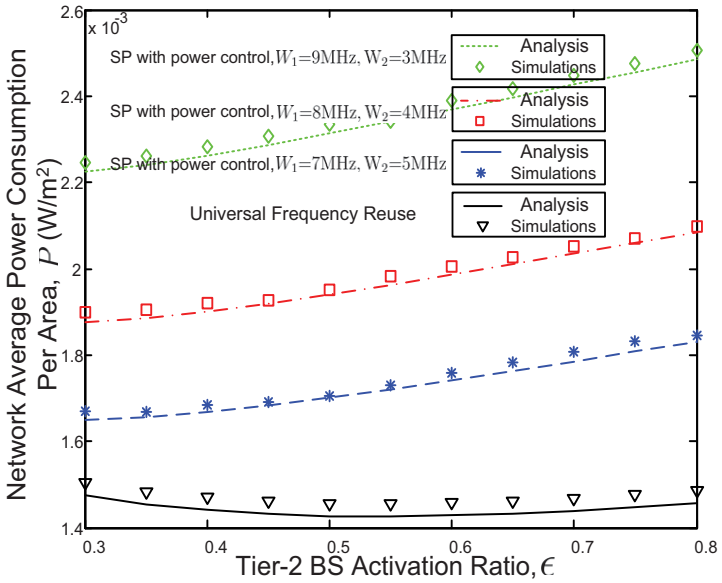


Figure 5.8: The network average power consumption per area P between universal frequency reuse and spectrum partitioning with power control. $\eta = 0.08$ and $\gamma\lambda_{\mathrm{u}}L = 17\mathrm{Mbps/km}^2$. For universal frequency reuse, $P_1 = 20\mathrm{W}$ and $P_2 = 6\mathrm{W}$. For spectrum partitioning with power control, P_1 and P_2 are obtained by solving (5.30)

coverage and the network mean queuing delay. It is shown that if the idle power coefficient is below a certain threshold, the optimal solution should equal the activation ratio to minimize the network average power consumption per area. Otherwise, the optimal activation ratio should be obtained according to the QoS constraints. Simulation results illustrate a significant improvement on the network energy efficiency by carefully choosing the activation ratio. It is further revealed that by taking queuing into account, universal spectrum reuse outperforms spectrum partitioning in terms of energy efficiency and SIR coverage in the considered scenario.

Chapter 6

Optimal Biased Association Scheme with Non-Uniform User Distribution

So far, Chapter 3, 4 and 5 have studied how to tune the optimal biasing factor of each tier, the bandwidth allocated to each tier, and the micro BS deployment density in HetNets, respectively, by taking queuing into consideration. Since the impact of a more practical characterization of the user distribution on the system performance still remains as an open problem, which is mentioned in Section 1.3, we will study the optimal biased association scheme with non-uniform user distribution in this chapter. In particular, in contrast to previous studies where users are usually assumed to be uniformly distributed, and thereby a per-tier SINR biasing factor is used to balance the load of BSs among different tiers, we examine in this chapter a scenario that one cell is overloaded, i.e., has a higher user intensity. In this case, the adjustment of the per-tier biasing factor becomes unreasonable, and thus we propose to adjust the biasing factor of the overloaded cell to offload the traffic to its surrounding cells. By maximizing the mean user utility in the area of this overloaded cell and its neighboring cells, the optimal biasing factor can be obtained. It is found that in the scenario where the overloaded cell is fully surrounded by

a macro cell, the optimal biasing factor logarithmically decreases with the user's intensity of the overloaded cell. Numerical results demonstrate that by using the optimal biasing factor of the overloaded cell in the considered scenario, both the mean user rate in the overloaded cell and the that of the whole network can be increased significantly compared to the traditional per-tier biased scheme without the adjustment of the overloaded cell in the literature. Our analysis in this chapter provides guidance on the optimal tuning of the biasing factor of an overloaded cell and, is a step forward towards the goal of the adjustment of the biasing factor in a per-station fashion under heterogeneous spatial user distribution.

6.1 Introduction

Recall that in HetNets, even with a targeted deployment where these small-scale BSs are placed in high-traffic zones, most users will still receive the strongest downlink signal from the tower-mounted macro BS, thus causing the load imbalance across tiers. Therefore, as a key component to realize the potential of capacity enhancement with the architecture of HetNets, biased association scheme has long been studied and attracted extensive attention [46, 66–68]. A detailed review can be found in Section 1.2.1.

However, the aforementioned studies [46, 66–68] assume homogeneous user distribution, i.e., users have a uniform population density. A per-tier biased scheme is thus adopted where BSs of the same tier share an identical biasing factor. In practice, however, users might not be evenly distributed. To be specific, users might form a cluster in some areas such as Hotspots. In such cases, the tuning of the biasing factor in a per-tier fashion would not relieve the traffic pressure in the overloaded areas, and a per-station biased scheme is thus preferable [30]. Although we follow a similar utility maximization approach in [68], this chapter examines a different scenario where one cell is overloaded

with a cluster of users, and thus the biasing factor of this particular cell is tuned based on its load condition. To the best of the our knowledge, this study is the first to propose a user intensity oriented biased scheme in the context of non-uniform user distribution in HetNets, and is a step forward towards the goal of tuning the biasing factor in a per-station fashion.

The contributions of this chapter are summarized as follows.

- A load-aware biased association scheme is proposed where the overloaded BS optimally adjusts its biasing factor according to its load condition by maximizing a utility function of the mean user rate;
- It is demonstrated that in a simple scenario where the overloaded cell is adjacent to only a macro cell, the optimal biasing factor of the overloaded BS can be perfectly fitted as a log-linear function of its user density of the overloaded cell. This observation greatly facilitates the implementation of the proposed scheme in practice;
- Simulation results show that although the performance of the macro cell deteriorates, the proposed scheme can significantly improve the mean user rate in the overloaded cell from 23% to 87% as the average number of cluster users increases from 10 to 80 compared to the previous biased scheme without the adjustment of the overloaded cell, and the overall mean user rate performance is also improved due to load balancing.

The rest of this chapter is organized as follows. System model is presented in Section 6.2. An optimization problem to maximize mean user utility is formulated in Section 6.3. A simple scenario is examined in Section 6.4. Simulations results are presented in Section 6.5. Conclusions and future works are given in Section 6.6.

6.2 System Model

6.2.1 Network Topology

Consider a K-Tier heterogeneous network with fixed locations of BSs. Each user in the network is associated with one of these BSs. We assume all the BSs share the spectrum. Therefore, for each user in the downlink, the associated BS acts as a desired signal transmitter while other BSs act as interfering sources. In contrast to previous studies where users are usually assumed to follow a homogeneous PPP with a uniform intensity [46, 66–68], we consider in this chapter that one cell has a different user density, which is a typical scenario in our daily life. For instance, the association region of a cell can be one hall or one room where people attend a lecture or enjoy a concert and thus form a cluster. As Fig. 6.1 illustrates, users in the whole area form a homogeneous PPP with the intensity λ_r while in one cell (the shaded area) there are additional users following another independent PPP with the intensity λ_c. As such, users in this cell form a superposition of two independent PPP with the intensity λ_r and λ_c, which mimics the case that this cell is overloaded when $\lambda_c > 0$. In the following, we refer to the users that follow a homogeneous distribution in the whole region as regular users, and those additional users inside the overloaded cell as cluster users. Denote this overloaded cell as $C_0 = \{\mathrm{BS}_0, A_0\}$ with a Tier-$k_0 \in \{1, \ldots, K\}$ base station BS_0 and its corresponding association region as A_0. Without loss of generality, we assume that Cell C_0 is adjacent to Cell $C_m = \{\mathrm{BS}_m, A_m\}$, $m = 1, \ldots, M$, where BS_m belongs to Tier $k_m \in \{1, \ldots, K\}$ and A_m is the corresponding association region. Note that C_0 is adjacent to C_m as long as $A_0 \cap A_m \neq \emptyset$[1].

[1] $A_0 \cap A_m$ is the boundary of two cells, and they are adjacent to each other if it is not a void set.

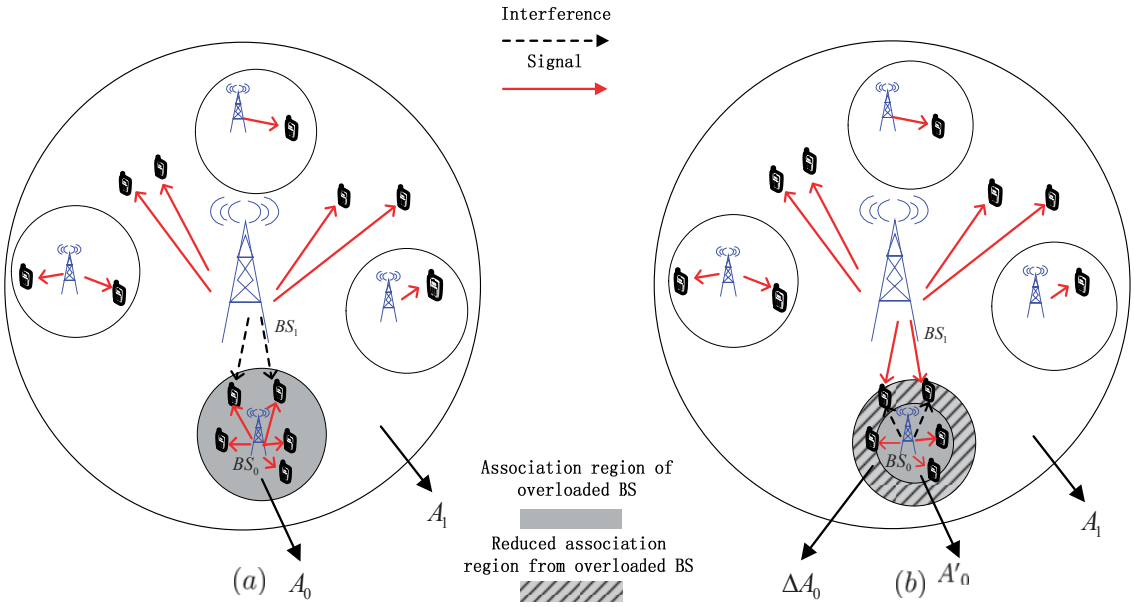


Figure 6.1: Illustration of the network topology. (a) Without adjustment of the biasing factor of the overloaded cell. (b) With adjustment of the biasing factor of the overloaded cell.

6.2.2 Association Region and Biasing Factor

For a random user, its received power P_R from a BS can be written as

$$P_R = P_S d_S^{-\alpha} g_S, \tag{6.1}$$

where P_S is the transmission power of the BS and d_S is the distance from this user to the BS. Note that α is the path loss coefficient which is assumed to be the same for all the BSs for simplicity, and g_S is the small-scale coefficient which is assumed to follow an independent and identical exponential distribution of unit mean, i.e., $g_S \sim \exp\{1\}$.

Without cluster users, i.e., $\lambda_c = 0$, all the users form a homogeneous PPP with the intensity λ_r. In this case, each user is associated to the BS with the maximum biased received power [59]. The tier of the associated BS is then given by

$$\arg \max_{k=1,\ldots,K} f_k P_k d_{k,\min}^{-\alpha}, \tag{6.2}$$

where f_k is the per-tier biasing factor reflecting association preference of a random user

towards a Tier-k BS, P_k is the transmission power of a Tier-k BS and $d_{k,\min}$ is the distance between the user and its nearest Tier-k BS.

When a cluster of users appear in the association region of Cell C_0, i.e., $\lambda_c > 0$, these users will also associate with BS_0 if the per-tier biasing factor is adopted. In this case, as the number of cluster users increases, each user would share a diminishing fraction of resources of the overloaded C_0. To improve the performance of these users, some of them should be offloaded to the neighboring cells. Intuitively, the per-tier bias may not be reasonable as the load of other BSs of the same tier does not change. Therefore, in this chapter we propose to tune the biasing factor of the overloaded BS based on its load condition.

In particular, we aim to characterize the biasing factor of Cell C_0, f'_{k_0}. It is clear that with $\lambda_c > 0$, BS_0 should decrease the biasing factor to offload its edge users to neighboring cells, thus we have $f'_{k_0} < f_{k_0}$. As for the M neighboring cells, the biasing factor remains the same, i.e., for Cell C_m where $m = 1, \ldots, M$, its biasing factor equals f_{k_m} where $k_m \in \{1, \ldots, K\}$. Hence, for each user in the region of the overloaded cell and its M neighboring cells, it would associate with BS_i, and the index i is given by

$$i = \arg \max_{m=0,\ldots,M} \beta_m, \tag{6.3}$$

where $\beta_m = \begin{cases} f_{k_m} P_{k_m} d_{k_m}^{-\alpha} & m = 1, ..., M \\ f'_{k_0} P_{k_0} d_{k_0}^{-\alpha} & m = 0 \end{cases}$.

Here P_{k_m} and d_{k_m} refer to the transmission power of BS_m and the distance of the user to that BS, respectively.

As $f'_{k_0} < f_{k_0}$, the association regions of the M neighboring cells expand while that of C_0 shrinks. In the following, we denote A_m and A'_m as the association regions with the biasing factor f_{k_0} and f'_{k_0}, respectively, and ΔA_m as the corresponding changed association region.

6.2.3 Mean User Rate

Similar to [68], we consider a fixed rate modulation and coding format, and assume that the interference dominates the background noise. In this case, a BS could serve a user only when the user's instantaneous SIR exceeds a threshold τ. By denoting I as the interference, the spectrum efficiency can be obtained as

$$\eta = \begin{cases} \log_2(1+\tau) & \text{SIR} \geqslant \tau \\ 0 & \text{SIR} < \tau \end{cases}, \tag{6.4}$$

where

$$\text{SIR} = \frac{P_S d_S^{-\alpha} g_S}{I} \tag{6.5}$$

according to (6.1). In addition, we adopt equal resource allocation, which is shown to maximize the proportional fairness [46]. By denoting N and W as the number of users associated to a BS and the total bandwidth, respectively, the spectrum allocated to each user that is associated to this BS can then be written as W/N. By combining (6.4) and (6.5), the mean rate of one user can be obtained as

$$\overline{R} = \frac{W}{N} \cdot \log_2(1+\tau) \Pr(\text{SIR} \geqslant \tau). \tag{6.6}$$

6.3 Mean User Utility Optimization

In this section we will derive the mean user utility. To strike a balance between throughput and fairness, the utility is defined as a logarithmic function of the mean user rate. By maximizing the mean user utility, we aim to find the optimal biasing factor for the overloaded BS. As mentioned in Section 6.2.2, the adjustment of the biasing factor for the overloaded BS only affects the association regions of this overloaded cell and its M neighboring cells. Therefore, we only need to consider the mean user utility concerning theses M+1 cells. Let $\Pr_A$ and $\overline{R}_A$ denote the probability and the mean

rate of a random user, respectively, given that this user is located in Region A, where $A \in \{A'_0, A_1, A_2, \ldots, A_M, \Delta A_1, \Delta A_2, \ldots, \Delta A_M\}$. The mean user utility can then be written as

$$U = \sum_{m=1}^{M} \left\{ \Pr_{A_m} E\left[\log\left(\overline{R}_{A_m}\right)\right] + \Pr_{\Delta A_m} E\left[\log\left(\overline{R}_{\Delta A_m}\right)\right] \right\} + \Pr_{A'_0} E\left[\log\left(\overline{R}_{A'_0}\right)\right]. \quad (6.7)$$

It is clear from (6.7) that the mean user utility is determined by the probability of a user being located in a Region A, $\Pr_A$, and the mean logarithm of the user rate in that region, $E\left[\log\left(\overline{R}_A\right)\right]$. In the following, we will derive these two components.

6.3.1 Probability of A Random User's Location

As mentioned in Section 6.2.1, regular users form a PPP in the region $\bigcup_{m=0}^{M} A_m$ with the intensity λ_r while cluster users form an independent PPP with the intensity λ_c. Denoting S_A as the association area of Region A, where $A \in \{A'_0, A_1, A_2, \ldots, A_M, \Delta A_1, \Delta A_2, \ldots, \Delta A_M\}$, the average number of regular users $\overline{N}_r$ and cluster users $\overline{N}_c$ can then be obtained as

$$\overline{N}_r = \lambda_r \sum_{m=0}^{M} S_{A_m} \quad (6.8a)$$

and

$$\overline{N}_c = \lambda_c S_{A_0}, \quad (6.8b)$$

respectively. The average number of all users $\overline{N}_t$ is thus given by

$$\overline{N}_t = \overline{N}_r + \overline{N}_c = \lambda_r \sum_{m=0}^{M} S_{A_m} + \lambda_c S_{A_0}. \quad (6.8c)$$

Denoting $\Pr_r$ and $\Pr_c$ as the probabilities that a random user is a regular user and a cluster user, respectively. We have, in the former case,

$$\Pr_r \approx \frac{\overline{N}_r}{\overline{N}_t} \quad (6.9a)$$

and in the latter case,

$$\Pr_c \approx \frac{\overline{N}_c}{\overline{N}_t}. \tag{6.9b}$$

By denoting **RU** and **CU** as the cases that a random user is a regular user and a cluster user, respectively, we have

$$\Pr_{A'_0} = \Pr_r \cdot \Pr\left(A'_0|\mathbf{RU}\right) + \Pr_c \cdot \Pr\left(A'_0|\mathbf{CU}\right) = \frac{\overline{N}_r S_{A'_0}}{\overline{N}_t \sum_{l=0}^{M} S_{A_l}} + \frac{\overline{N}_c S_{A'_0}}{\overline{N}_t S_{A_0}}, \tag{6.10a}$$

$$\Pr_{\Delta A_m} = \Pr_r \cdot \Pr\left(\Delta A_m|\mathbf{RU}\right) + \Pr_c \cdot \Pr\left(\Delta A_m|\mathbf{CU}\right) = \frac{\overline{N}_r S_{\Delta A_m}}{\overline{N}_t \sum_{l=0}^{M} S_{A_l}} + \frac{\overline{N}_c S_{\Delta A_m}}{\overline{N}_t S_{A_0}}, \tag{6.10b}$$

and

$$\Pr_{A_m} = \Pr_r \cdot \Pr\left(A_m|\mathbf{RU}\right) = \frac{\overline{N}_r S_{A_m}}{\overline{N}_t \sum_{l=0}^{M} S_{A_l}}, \tag{6.10c}$$

where $m = 1, \ldots, M$, according to (6.8) and (6.9).

6.3.2 Mean Logarithm of User Rate

According to (6.6), we have

$$E\left[\log\left(\overline{R}_A\right)\right] = \log\left(W\log_2\left(1+\tau\right)\right) + E\left[\log\left(\Pr\left(\mathrm{SIR}_A \geqslant \tau\right)\right)\right] - E\left[\log\left(N_m\right)\right] \tag{6.11}$$

where SIR_A denotes the signal to noise ratio for a random user in Region A, and N_m is the total number of users associated to $\mathrm{BS}_m, m \in \{0, \ldots, M\}$. Similar to [68], the mean logarithm of BS_m's load, $E\left[\log\left(N_m\right)\right]$, can be approximated as its upper bound $\log\left(E\left[N_m\right]\right)$. We then have

$$E\left[\log\left(N_0\right)\right] \leqslant \log\left(E\left[N_0\right]\right) = \log\left(\left(\lambda_c + \lambda_r\right) S_{A'_0}\right) \tag{6.12a}$$

and

$$E\left[\log\left(N_m\right)\right] \leqslant \log\left(E\left[N_m\right]\right) = \log\left(\lambda_r\left(S_{A_m} + S_{\Delta A_m}\right) + \lambda_c S_{\Delta A_m}\right), \tag{6.12b}$$

where $m = 1, \ldots, M$.

Finally, by substituting (6.10), (6.11) and (6.12) into (6.7), we can obtain the mean user utility in the overloaded cell and its neighboring M cells. The optimal biasing factor of the overloaded BS, f^*, can then be obtained by maximizing the mean user utility,

$$\begin{aligned} f^* \overset{(a)}{=} \arg\max_{f'_{k_0}} & -\sum_{m=1}^{M} \frac{\overline{N}_r S_{A_m}}{\overline{N}_t \sum_{l=0}^{M} S_{A_l}} \cdot \log\left(\frac{\overline{N}_r\left(S_{A_m} + S_{\Delta A_m}\right)}{\sum_{m=0}^{M} S_{A_m}} + \frac{\overline{N}_c S_{\Delta A_m}}{S_{A_0}}\right) \\ & + \sum_{m=1}^{M}\left(\frac{\overline{N}_r S_{\Delta A_m}}{\overline{N}_t \sum_{l=0}^{M} S_{A_l}} + \frac{\overline{N}_c S_{\Delta A_m}}{\overline{N}_t S_{A_0}}\right) \cdot \Bigg\{ C + E\left[\log\left(\Pr\left(\mathrm{SIR}_{\Delta A_m} \geqslant \tau\right)\right)\right] \\ & - \log\left(\frac{\overline{N}_r\left(S_{A_m} + S_{\Delta A_m}\right)}{\sum_{m=0}^{M} S_{A_m}} + \frac{\overline{N}_c S_{\Delta A_m}}{S_{A_0}}\right)\Bigg\} + \left(\frac{\overline{N}_r S_{A'_0}}{\overline{N}_t \sum_{l=0}^{M} S_{A_l}} + \frac{\overline{N}_c S_{A'_0}}{\overline{N}_t S_{A_0}}\right) \\ & \cdot \Bigg\{ C + E\left[\log\left(\Pr\left(\mathrm{SIR}_{A'_0} \geqslant \tau\right)\right)\right] - \log\left(\left(\frac{\overline{N}_c}{S_{A_0}} + \frac{\overline{N}_r}{\sum_{m=0}^{M} S_{A_m}}\right) S_{A'_0}\right)\Bigg\}, \end{aligned} \tag{6.13}$$

where $C = \log\left(W \log_2\left(1+\tau\right)\right)$ and (a) follows the fact that optimal biasing factor f'_{k_0} does not depend on $\Pr_{A_m} E\left[\log\left(\overline{R}_{A_m}\right)\right]$.

To this end, we have presented the proposed scheme, which is summarized in the flowchart in Fig. 6.2. As illustrated in Fig. 6.2, for the base station BS_0, it firstly measures the associated user intensity within the corresponding association region A_0, and exchange its intensity information with neighboring BSs. After obtaining the intensity of regular users λ_r and that of cluster users λ_c, it can decide if there exists a cluster within its region.

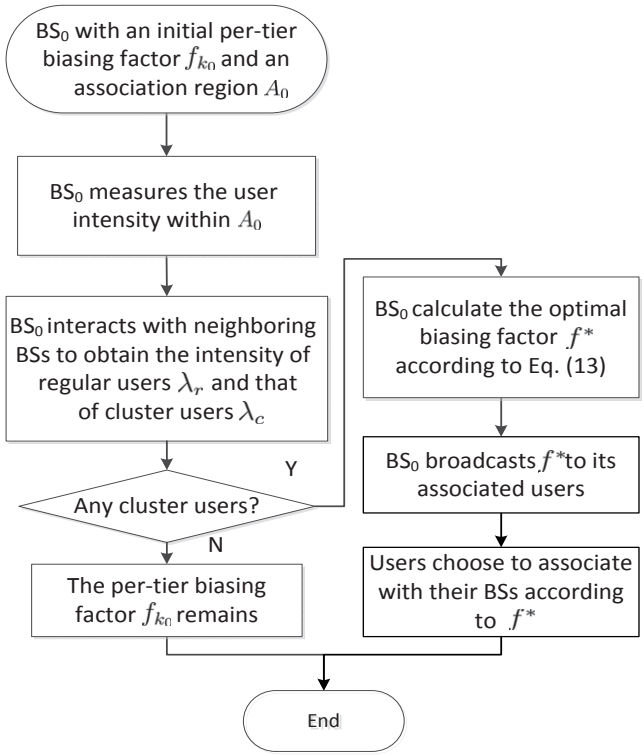


Figure 6.2: Flowchart of the proposed scheme.

If there are no cluster users, BS_0 keeps the per-tier biasing factor f_{k_0}. Otherwise, BS_0 would calculate the optimal biasing factor f^* according to (6.13), and then broadcast to its associated users. Upon receiving the optimal biasing factor f^*, each user would choose a new serving BS accordingly.

As we can see from (6.13), the mean user utility depends on the network topology. In the following we will illustrate the above analytical results by examining a simple network scenario.

6.4 Case Study

We consider in this section that the overloaded cell C_0 is adjacent to only one macro cell C_1, i.e., $A_0 \cap A_1 \neq \emptyset$. As illustrated in Fig. 6.1, C_0 is fully surrounded by the macro cell C_1 and is not adjacent to other cells. If more users emerge in the association area of C_0,

i.e., as λ_c increases, the biasing factor of Cell C_0 should be properly reduced and edge users in A_0 will be pushed to the macro cell C_1, which is demonstrated in Fig. 1(b). This scenario corresponds to the situation that low-tier BSs do not provide open access. For example, privately-owned access points only support connection requests from authorized users. In this case, users in the overloaded micro cell can only be offloaded to the macro cell.

For each user in Region A_0', the received power from the BS_1 is much higher than that from other BSs; therefore, the interference in Region A_0' can be written as $I_{A_0'} \approx P_{k_1} d_{k_1}^{-\alpha} g_{k_1}$. On the other hand, for each user in Region ΔA_0 which becomes part of the association region of BS_1 after the adjustment of the biasing factor of BS_0, the received power from BS_0 is the main source of the interference. We then have $I_{\Delta A_0} \approx P_{k_0} d_{k_0}^{-\alpha} g_{k_0}$.

According to [66], by denoting the location of BS_0, BS_1 and a random user as $\mathbf{l_0}$, $\mathbf{l_1}$ and $\mathbf{l_u}$, respectively, the association region of C_0 can be obtained as

$$A_0 = \left\{ \mathbf{l_u} : P_{k_1} \|\mathbf{l_u} - \mathbf{l_1}\|^{-\alpha} \leqslant \bar{f}_{k_0} P_{k_0} \|\mathbf{l_u} - \mathbf{l_0}\|^{-\alpha} \right\}, \tag{6.14}$$

where

$$\bar{f}_{k_0} = f_{k_0} / f_{k_1} \tag{6.15}$$

is the normalized biasing factor of BS_0 with respect to BS_1, and $\|\cdot\|$ indicates the Euclidean distance. By letting $\mathbf{l_1} = (0, 0)$ and $\mathbf{l_0} = (a, 0)$, (6.14) can be further written as

$$A_0 = \left\{ (x, y) : \left(x - \frac{Ba}{B-1} \right)^2 + y^2 \leqslant \frac{Ba^2}{(B-1)^2} \right\}, \tag{6.16a}$$

where $B = \left(\frac{P_{k_1}}{P_{k_0} \bar{f}_{k_0}} \right)^{\frac{2}{\alpha}}$. It is clear from (6.16a) that A_0 is a circle with the center $\left(\frac{Ba}{B-1}, 0 \right)$ and the radius $r = \sqrt{\frac{Ba^2}{(B-1)^2}}$. Similarly, by denoting $\bar{f}_{k_0}' = f_{k_0}' / f_{k_1}$ as the adjusted normal-

ized biasing factor of BS_0, A'_0 can be obtained as

$$\begin{aligned} A'_0 &= \left\{ \mathbf{l_u} : P_{k_1} \|\mathbf{l_u} - \mathbf{l_1}\|^{-\alpha} \leqslant \bar{f}'_{k_0} P_{k_0} \|\mathbf{l_u} - \mathbf{l_0}\|^{-\alpha} \right\} \\ &= \left\{ (x, y) : \left(x - \frac{B'a}{B'-1} \right)^2 + y^2 \leqslant \frac{B'a^2}{(B'-1)^2} \right\}, \end{aligned} \tag{6.16b}$$

which is also a circle with the center $\left(\frac{B'a}{B'-1}, 0\right)$ and the radius $r' = \sqrt{\frac{B'a^2}{(B'-1)^2}}$, where $B' = \left(\frac{P_{k_1}}{P_{k_0}\bar{f}'_{k_0}}\right)^{\frac{2}{\alpha}}$.

Accordingly, we have

$$S_{A_0} = \pi r^2, \tag{6.17a}$$

$$S_{A'_0} = \pi r'^2, \tag{6.17b}$$

and

$$S_{\Delta A_0} = \pi \left(r^2 - r'^2 \right). \tag{6.17c}$$

By substituting (6.17) into (6.8), we can obtain that

$$\overline{N}_r = \lambda_r \left(\pi r^2 + S_{A_1} \right), \tag{6.18a}$$

$$\overline{N}_c = \lambda_c \pi r^2, \tag{6.18b}$$

and

$$\overline{N}_t = \lambda_r \left(\pi r^2 + S_{A_1} \right) + \lambda_c \pi r^2, \tag{6.18c}$$

where S_{A_1} can be calculated by subtracting other cell's association regions from that of the macro cell. By combining (6.10), (6.12), (6.17) and (6.18), we have

$$\mathrm{Pr}_{A'_0} = \frac{\overline{N}_r \pi r'^2}{\overline{N}_t \left(\pi r^2 + S_{A_1} \right)} + \frac{\overline{N}_c r'^2}{\overline{N}_t r^2}, \tag{6.19a}$$

$$\Pr_{A_1} = \frac{\overline{N}_r \pi r^2}{\overline{N}_t \left(\pi r^2 + S_{A_1}\right)}, \tag{6.19b}$$

$$\Pr_{\Delta A_0} = \frac{\overline{N}_r \pi \left(r^2 - r'^2\right)}{\overline{N}_t \left(\pi r^2 + S_{A_1}\right)} + \frac{\overline{N}_c \left(r^2 - r'^2\right)}{\overline{N}_t r^2}, \tag{6.19c}$$

and

$$E\left[\log\left(N_0\right)\right] \approx \log\left(\left(\lambda_c + \lambda_r\right)\pi r'^2\right), \tag{6.20a}$$

$$E\left[\log\left(N_1\right)\right] \approx \log\left(\lambda_r S_{A_1} + \left(\lambda_r + \lambda_c\right)\pi\left(r^2 - r'^2\right)\right). \tag{6.20b}$$

Furthermore, it can be obtained that

$$E\left[\log\left(\Pr\left(\mathrm{SIR}_{A'_0} \geqslant \tau\right)\right)\right] = -\oiint_{A'_0} \left(\frac{\left(x-a\right)^2 + y^2}{x^2 + y^2}\right)^{\frac{\alpha}{2}} \frac{\tau P_{k_1} P_{k_0}^{-1}}{\pi r'^2} dxdy \tag{6.21a}$$

and

$$E\left[\log\left(\Pr\left(\mathrm{SIR}_{\Delta A_0} \geqslant \tau\right)\right)\right] = -\oiint_{\Delta A_0} \left(\frac{x^2 + y^2}{\left(x-a\right)^2 + y^2}\right)^{\frac{\alpha}{2}} \frac{\tau P_{k_0} P_{k_1}^{-1}}{\pi\left(r^2 - r'^2\right)} dxdy. \tag{6.21b}$$

The proof of (6.21) can be found in the following.

Proof. For a random user in Region A'_0 where the undesired signal from BS_1 is the major source of the interference, we have

$$E\left[\log\left(\Pr\left(\mathrm{SIR}_{A'_0} \geqslant \tau\right)\right)\right] = \int_0^\infty E_{d_{k_0}, g_{k_0}, d_{k_1}}\left[\log\left(\Pr\left(\frac{P_{k_0} d_{k_0}^{-\alpha} g_{k_0}}{P_{k_1} d_{k_1}^{-\alpha} g_{k_1}} \geqslant \tau\right)\right) | g_{k_1}\right] \cdot pdf_{g_{k_1}}\left(g_{k_1}\right) dg_{k_1}, \tag{6.22}$$

where $pdf_{g_{k_1}}\left(g_{k_1}\right)$ is probability density function of random variable g_{k_1}. Conditioned on a given g_{k_1} we have

$$E_{d_{k_0}, g_{k_0}, d_{k_1}}\left[\log\left(\Pr\left(\frac{P_{k_0} d_{k_0}^{-\alpha} g_{k_0}}{P_{k_1} d_{k_1}^{-\alpha} g_{k_1}} \geqslant \tau\right)\right) | g_{k_1}\right] =$$
$$\oiint E_{g_{k_0}}\left[\log\left(\Pr\left(\frac{P_{k_0} d_{k_0}^{-\alpha} g_{k_0}}{P_{k_1} d_{k_1}^{-\alpha} g_{k_1}} \geqslant \tau\right)\right) | g_{k_1}, d_{k_0}, d_{k_1}\right] \cdot pdf_{d_{k_0}, d_{k_1}}\left(d_{k_0}, d_{k_1}\right) dd_{k_0} dd_{k_1}, \tag{6.23}$$

where $pdf_{d_{k_0},d_{k_1}}(d_{k_0},d_{k_1})$ denotes the joint PDF of d_{k_0} and d_{k_1}. Hence, conditioned on g_{k_1}, d_{k_0} and d_{k_1}, we can obtain that

$$\begin{aligned}
&E_{g_{k_0}}\left[\log\left(\Pr\left(\frac{P_{k_0}d_{k_0}^{-\alpha}g_{k_0}}{P_{k_1}d_{k_1}^{-\alpha}g_{k_1}}\geqslant\tau\right)\right)|g_{k_1},d_{k_0},d_{k_1}\right]\\
&\overset{(a)}{=}E_{g_{k_0}}\left[\log\left(\exp\left(-\tau d_{k_1}^{-\alpha}d_{k_0}^{\alpha}P_{k_1}P_{k_0}^{-1}\right)\right)|g_{k_1},d_{k_0},d_{k_1}\right]\\
&=-\tau d_{k_1}^{-\alpha}d_{k_0}^{\alpha}P_{k_1}P_{k_0}^{-1}g_{k_1},
\end{aligned}\tag{6.24}$$

where (a) follows that g_{k_0} is an exponential random variable with unit mean.

By substituting (6.23) and (6.24) into (6.22), we have

$$\begin{aligned}
&E\left[\log\left(\Pr\left(\mathrm{SIR}_{A_0'}\geqslant\tau\right)\right)\right]\\
&=-\tau P_{k_1}P_{k_0}^{-1}\int_0^{\infty}g_{k_1}pdf_{g_{k_1}}(g_{k_1})\,dg_{k_1}\cdot\oiint\frac{d_{k_0}^{\alpha}}{d_{k_1}^{\alpha}}pdf_{d_{k_0},d_1}(d_{k_0},d_{k_1})\,dd_{k_0}dd_{k_1}\\
&\overset{(b)}{=}-\tau P_{k_1}P_{k_0}^{-1}\oiint_{A_0'}\left(\frac{(x-a)^2+y^2}{x^2+y^2}\right)^{\frac{\alpha}{2}}pdf_{x,y}(x,y)\,dxdy\\
&\overset{(c)}{=}-\oiint_{A_0'}\left(\frac{(x-a)^2+y^2}{x^2+y^2}\right)^{\frac{\alpha}{2}}\frac{\tau P_{k_1}P_{k_0}^{-1}}{\pi r'^2}dxdy,
\end{aligned}\tag{6.25}$$

where (b) follows that g_{k_1} is exponentially distributed with unit mean, i.e., $E_{g_{k_1}}[g_{k_1}]=1$, and (c) follows that users are uniformly deployed within each association region of a cell, i.e., $pdf_{x,y}(x,y)=1/\pi r'^2$.

On the other hand, $E\left[\log\left(\Pr\left(\mathrm{SIR}_{\Delta A_0}\geqslant\tau\right)\right)\right]$ can be derived by a similar approach and thus omitted here. □

Finally, by substituting (6.8), (6.18), (6.19), (6.20) and (6.21) into (6.13), the normalized optimal biasing factor for the overloaded C_0 can be obtained as

$$f^*=\arg\max_{\bar{f}_{k_0}'}O,\tag{6.26}$$

where the objective function

$$
\begin{aligned}
O = &- \frac{\overline{N}_r \pi r^2}{\overline{N}_t \left(\pi r^2 + S_{A_1}\right)} \cdot \log\left(\frac{\overline{N}_r S_{A_1}}{\pi r^2 + S_{A_1}} + \left(\frac{\overline{N}_c}{\pi r^2} + \frac{\overline{N}_r}{\pi r^2 + S_{A_1}}\right)\pi\left(r^2 - r'^2\right)\right) + \\
&\left(\frac{\overline{N}_r \pi \left(r^2 - r'^2\right)}{\overline{N}_t \left(\pi r^2 + S_{A_1}\right)} + \frac{\overline{N}_c \left(r^2 - r'^2\right)}{\overline{N}_t r^2}\right) \cdot \left(C - \oiint_{\Delta A_0} \left(\frac{x^2+y^2}{(x-a)^2+y^2}\right)^{\frac{\alpha}{2}} \cdot \right. \\
&\left. \frac{\tau P_{k_0} P_{k_1}^{-1}}{\pi\left(r^2 - r'^2\right)} dxdy - \log\left(\frac{\overline{N}_r S_{A_1}}{\pi r^2 + S_{A_1}} + \left(\frac{\overline{N}_c}{\pi r^2} + \frac{\overline{N}_r}{\pi r^2 + S_{A_1}}\right)\pi\left(r^2 - r'^2\right)\right)\right) \\
&+ \left(\frac{\overline{N}_r \pi r'^2}{\overline{N}_t \left(\pi r^2 + S_{A_1}\right)} + \frac{\overline{N}_c r'^2}{\overline{N}_t r^2}\right) \cdot \left(C - \oiint_{A_0'} \left(\frac{(x-a)^2+y^2}{x^2+y^2}\right)^{\frac{\alpha}{2}} \frac{\tau P_{k_1} P_{k_0}^{-1}}{\pi r'^2} dxdy \right. \\
&\left. - \log\left(\left(\frac{\overline{N}_c}{\pi r^2} + \frac{\overline{N}_r}{\pi r^2 + S_{A_1}}\right)\pi r'^2\right)\right)
\end{aligned} \tag{6.27}
$$

and the radius r' is given by

$$
r' = \frac{a\left(P_{k_1} / \left(P_{k_0} \bar{f}'_{k_0}\right)\right)^{\frac{1}{\alpha}}}{\left(P_{k_1} / \left(P_{k_0} \bar{f}'_{k_0}\right)\right)^{\frac{2}{\alpha}} - 1}. \tag{6.28}
$$

Even for this simple case, it is difficult to obtain a closed-form solution of the optimization problem (6.26), and f^* can only be obtained numerically. It will nevertheless be shown in the following section that f^* can be perfectly fitted as a log-linear function of the average number of cluster users given the power difference between the macro and the overloaded micro BS and average number of regular users.

Remark: If there is a large difference in the transmission power between BS_0 and BS_1, i.e., $P_{k_1} \gg P_{k_0}$, the objective function can be further written as

$$
\begin{aligned}
O' = &\frac{\overline{N}_c r'^2}{\overline{N}_t r^2}\left(C - \tau \frac{2P_{k_1}}{(\alpha+2)P_{k_0}}\left(\frac{r'}{a}\right)^{\alpha} - \log\left(\frac{\overline{N}_c r'^2}{r^2}\right)\right) + \\
&\frac{\overline{N}_c \left(r^2 - r'^2\right)}{\overline{N}_t r^2}\left(C - \tau \frac{P_{k_0}}{P_{k_1}} \frac{2a^{\alpha}}{(r-r')^2}\left(\frac{r^{2-\alpha} - r'^{2-\alpha}}{2-\alpha} - \frac{r' r^{1-\alpha} - r'^{2-\alpha}}{1-\alpha}\right)\right) \\
&- \left(\frac{\overline{N}_c \left(r^2 - r'^2\right)}{\overline{N}_t r^2} + \frac{\overline{N}_r}{\overline{N}_t}\right) \log\left(\frac{\overline{N}_r \pi r'^2}{\pi r^2 + S_{A_1}} + \frac{\overline{N}_c \left(r^2 - r'^2\right)}{r^2}\right).
\end{aligned} \tag{6.29}
$$

Proof. With $P_{k_1} \gg P_{k_0}$, the radius of Region A_0 is much smaller than the distance between BS_1 and the overloaded BS_0, i.e., $r = \sqrt{\frac{Ba^2}{(B-1)^2}} \ll a$. We then have $B \gg 1$, and thus $\frac{Ba}{B-1} \approx a$. In addition, as the biasing factor of BS_0 decreases to offload some users to BS_1, we have $B' > B \gg 1$, and thus $\frac{Ba}{B-1} = \frac{B'a}{B'-1} \approx a$. Therefore, both A_0 and A_0' can be regarded as circular areas with the same center $(a, 0)$ but different radius. Meanwhile, since $r' < r \ll a$, the distance from BS_1 to a random user in region A_0 or A_0' can be approximately regarded as a constant, i.e., $d_{k_1} \approx a$. We then have

$$\begin{aligned} E\left[\log\left(\Pr\left(\mathrm{SIR}_{A_0'} \geqslant \tau\right)\right)\right] &\approx -\tau \frac{P_{k_1}}{P_{k_0}} E_{d_{k_0}, d_{k_1}}\left[\frac{d_{k_0}^{\alpha}}{a^{\alpha}}\right] \\ &\overset{(a)}{=} -\tau \frac{P_{k_1}}{P_{k_0} a^{\alpha}} \int_0^{r'} d_{k_0}^{\alpha} \cdot \frac{2 d_{k_0}}{r'^2} d d_{k_0} = -\frac{2\tau P_{k_1}}{(\alpha+2)\, P_{k_0}} \left(\frac{r'}{a}\right)^{\alpha} \end{aligned} \tag{6.30a}$$

and

$$\begin{aligned} E\left[\log\left(\Pr\left(\mathrm{SIR}_{\Delta A_0} \geqslant \tau\right)\right)\right] &\approx -\tau \frac{P_{k_0}}{P_{k_1}} E_{d_{k_0}, d_{k_1}}\left[\frac{a^{\alpha}}{d_{k_0}^{\alpha}}\right] \\ &\overset{(b)}{=} -\tau \frac{P_{k_0}}{P_{k_1}} a^{\alpha} \int_{r'}^{r} \frac{1}{d_{k_0}^{\alpha}} \frac{2\left(d_{k_0} - r'\right)}{\left(r - r'\right)^2} d d_{k_0} \\ &= C - \frac{2\tau P_{k_0} a^{\alpha}}{P_{k_1} (r - r')^2} \left(\frac{r^{2-\alpha} - r'^{2-\alpha}}{2-\alpha} - \frac{r' r^{1-\alpha} - r'^{2-\alpha}}{1-\alpha}\right), \end{aligned} \tag{6.30b}$$

where (a) and (b) follow the property of uniform distribution of a random user inside a circular area.

As regular users' intensity is usually much lower than that of cluster users, i.e., $\lambda_r \ll \lambda_c$, and the association area of the macro cell is much larger than that of other cells, (6.19) and (6.20) can be further written as

$$\Pr_{A_0'} = \frac{\overline{N}_c r'^2}{\overline{N}_t r^2}, \tag{6.31a}$$

$$\Pr_{A_1} = \frac{\overline{N}_r}{\overline{N}_t}, \tag{6.31b}$$

$$\Pr_{\Delta A_0} = \frac{\overline{N}_c (r - r')^2}{\overline{N}_t r^2}, \tag{6.31c}$$

Table 6.1: Simulation Parameters

Parameter	Value
Radius R	300 m
Distance a	150 m
Path Loss Coefficient α	4
Bandwidth BW	1 MHz
Noise Power σ_n^2	-104 dBm
SINR Threshold τ	2

and

$$E\left[\log\left(N_0\right)\right] \approx \log\left(\lambda_c \pi {r'}^2\right), \tag{6.32a}$$

$$E\left[\log\left(N_1\right)\right] \approx \log\left(\lambda_r S_{A_1} + \lambda_c \pi (r - r')^2\right). \tag{6.32b}$$

Finally, by substituting (6.18), (6.30), (6.31) and (6.32) into (6.13), the objective function (6.29) can be obtained. □

It can be proved that objective function (6.29) is concave with respect to r', by showing that $\frac{\partial^2 O'}{\partial {r'}^2} < 0$. Hence, the normalized optimal biasing factor can be obtained by combining $\frac{\partial O'}{\partial r'} = 0$ and (6.28).

6.5 Simulation Results

In this section we present simulation results to validate the analytical results in Section 6.4. Without loss of generality, we consider the whole region as a circular area with a radius of 300 meters. One Tier-1 base station BS_1 is located at the center $(0, 0)$. Four Tier-2 BSs scatter around BS_1 located at $(0, \pm a)$ and $(\pm a, 0)$, respectively. One of the four Tier-2 base stations, BS_0, is overloaded by a cluster of users with the intensity λ_c.

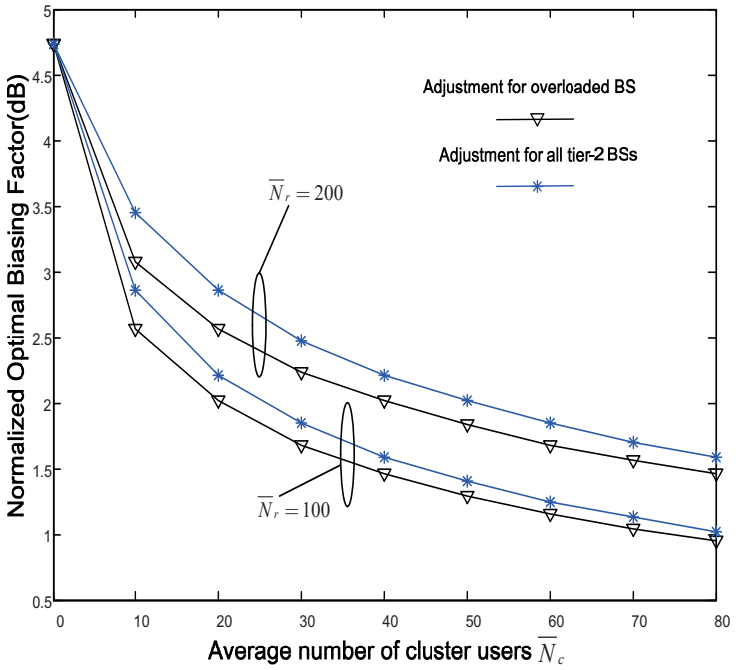


Figure 6.3: Normalized optimal biasing factor of the overloaded BS and all Tier-2 BSs with different values of average number of regular users, $\overline{N}_r$. $P_{k_1} = 60$dBm, $P_{k_0} = 20$dBm and $r = 20$m.

The normalized biasing factor f_{k_0} with $\lambda_c = 0$ determines the initial radius r of the Tier-2 cells. As cluster users appear in BS_0 with the initial radius r, the average number of cluster users is given by $\overline{N}_c = \lambda_c \pi r^2$. The system parameters are summarized in Table 6.1.

Fig. 6.3 compares the normalized optimal biasing factor of the overloaded BS_0 with that of all Tier-2 BSs. Note that the optimal per-tier biasing factor can be obtained by following a similar approach in this chapter, and is omitted here. As Fig. 6.3 illustrates, the optimal biasing factor of the overloaded BS_0 is smaller than that obtained by a per-tier adjustment. Intuitively, as the per-tier adjustment will cause all Tier-2 BSs to reduce their association regions, the resulting optimal biasing factor should be larger than f^* to keep these lightly-loaded micro cells more attractive to regular users. It can also be observed from Fig. 6.3 that both the normalized optimal biasing factors decrease as the

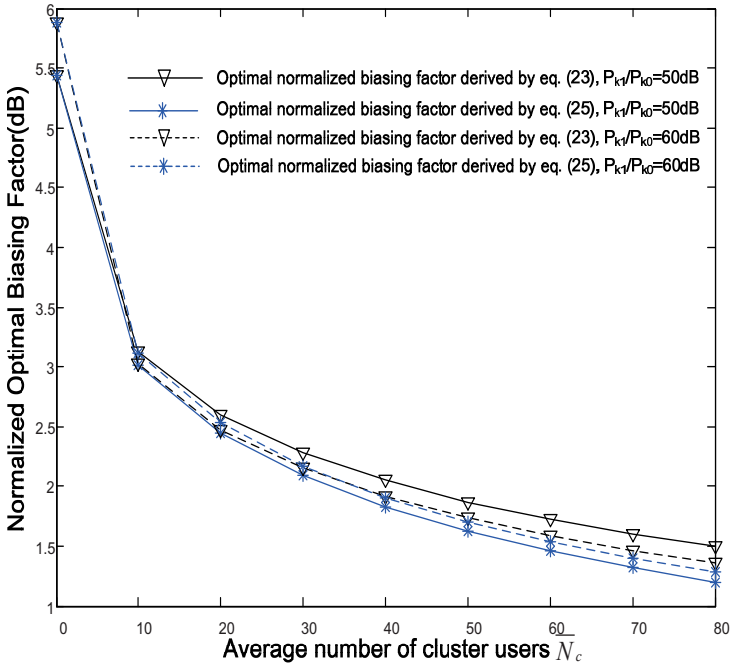


Figure 6.4: Normalized optimal biasing factor of the overloaded BS_0. $\overline{N}_r = 200$. $P_{k_1}/P_{k_0} = 50$dB corresponds to $r = 11.6$m and $P_{k_1}/P_{k_0} = 60$dB corresponds to $r = 6.7$m.

average number of cluster users increases, indicating that BS_0 should push out more users by shrinking its association region. Furthermore, the declining rate of the optimal factor f^* reduces as $\overline{N}_c$ becomes larger. This is because as BS_0 gets more crowded, users who are nearer to BS_0 would be offloaded to the macro BS, i.e., BS_1. As a result, the SIR of these users would drop more seriously, which indicates BS_1 should be less preferable. In addition, as more users choose to associate with BS_1, it should become less attractive to the cluster users located closer to the overloaded BS_0.

Fig. 6.4 demonstrates the curves of the normalized optimal biasing factor for the overloaded BS_0. In the remark of Section 6.4, when the transmission power of BS_1 is much larger than that of BS_0, the objective function given in (6.27) can be approximately written as (6.29). It can be observed from Fig. 6.4 that the gap between the normalized optimal biasing factors obtained from (6.27) and (6.29) indeed diminishes as the difference of the transmission power between BS_1 and BS_0 becomes larger.

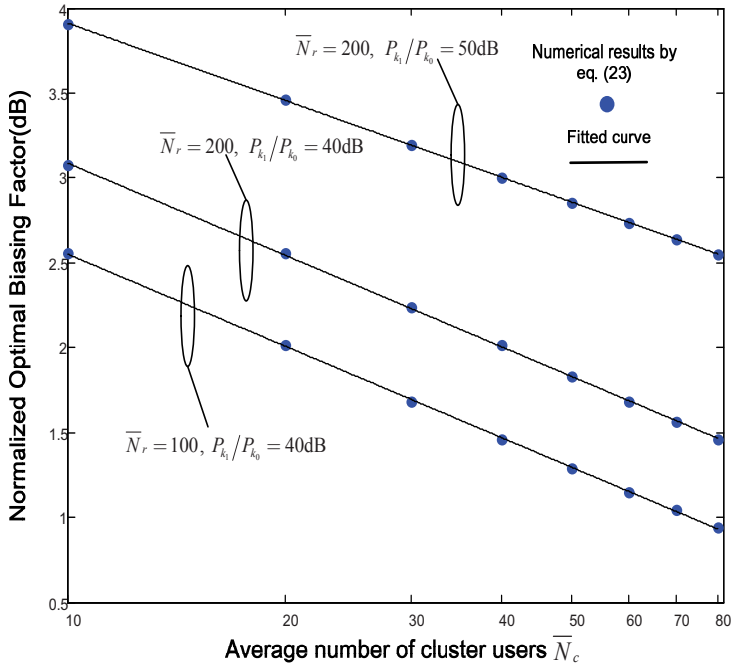


Figure 6.5: Numerical results and fitted curves of normalized optimal biasing factor of the overloaded BS. $r = 20$m. $P_{k_1}/P_{k_0} = 40$dB corresponds to the initial biasing factor of 4.73dB and $P_{k_1}/P_{k_0} = 50$dB corresponds to the initial biasing factor of 5.44dB

Fig. 6.5 illustrates how the optimal normalized biasing factor f^* varies with the average number of cluster users, $\overline{N}_c$ under various values of power difference between BS_0 and BS_1 and the average number of regular users. It can be clearly seen from Fig. 6.5 that f^* can be perfectly fitted into a log-linear function with the form $f^* = a_1 + a_2 \cdot \log_{10}\left(\overline{N}_c\right)$ when $\overline{N}_c$ is larger than 10. Although it is indicated in (6.27) that the optimal normalized biasing factor f^* is determined by the power difference P_{k_1}/P_{k_0}, the average number of regular users $\overline{N}_r$ and that of cluster users $\overline{N}_c$, it can be observed from Fig. 6.5 that the slope a_2 is not sensitive to these parameters. The intercept a_1, on the other hand, depends on the average number of regular users, $\overline{N}_r$. A larger $\overline{N}_r$ leads to a lower f^* as BS_1 needs to be more attractive if there are more regular users.

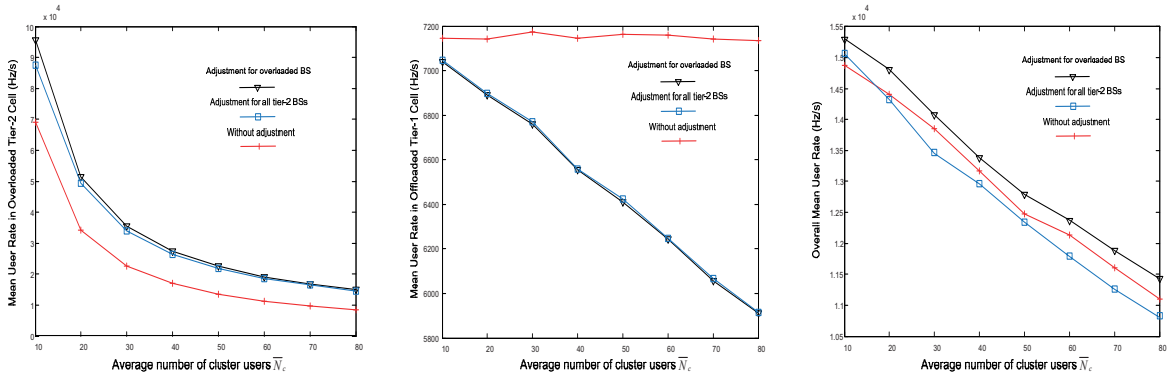
The results shown in Fig. 6.5 greatly facilitates the network design in practice. In particular, the two coefficients a_1 and a_2 are fixed given the power difference and the

average number of regular users. As cluster users would initially associate with a BS, this overloaded BS can estimate its average cluster numbers. The optimal biased value can then be obtained according to the log-linear function.

For general cases, f^* may not be a log-linear function of $\overline{N}_c$. The optimal biasing factor of the overloaded BS should instead be obtained through a linear search of (6.13). The computational complexity given the intensity of cluster users is then given by f_{k_0}/Δ, where f_{k_0} is the original biasing factor of the overloaded cell and Δ is searching step size. Note that it has been shown in Section 6.3 that with fixed intensity of regular users and BSs' transmitting powers, the optimal normalized biasing factor is solely determined by the intensity of cluster users. Hence, this mapping relation between f^* and λ_c can be numerically obtained by solving (6.13), and stored in the form of a table beforehand. The computational complexity for the table is $(f_{k_0}L)/\Delta$, where L could be the maximum possible number of cluster users in one specific practical scenario. As long as λ_c remains unchanged, the overloaded BS can obtain the optimal normalized biasing factor by simply looking up the table.

In practice, nevertheless, the intensity of regular users may vary with time. The table thus needs to be updated accordingly. By denoting φ_r as the changing frequency of λ_r, the computational complexity per unit time is thus $(f_{k_0}L\varphi_r)/\Delta$. In the considered scenario in this chapter, the intensity of regular users is assumed to be stable for a long period, compared to the changing frequency of the intensity of cluster users, i.e., φ_r is small. In this case, the computational complexity per unit time is quite low, indicating a good scalability of the proposed scheme.

Fig. 6.6 further compares the overall mean user rate, the mean user rate of the overloaded Tier-2 micro cell and the Tier-1 macro cell when only the biasing factor of the overloaded BS_0 is optimally tuned with the corresponding rates when the biasing factor of all Tier-2 BSs is optimally tuned. Each point is obtained by averaging over 2000 trials. It



(a) (b) (c)

Figure 6.6: Mean user rate in different cells. $\overline{N}_r = 200$, $P_{k_1} = 60$dBm, $P_{k_0} = 20$dBm, $r = 20$m. (a) Rates in the overloaded tier-2 cell. (b) Rates in the tier-1 macro cell. (c) Overall rates.

can be observed that compared to the previous biased scheme without tuning the biasing factor of BS_0, our proposed scheme can remarkably improve the mean user rate in such overloaded cell from 23% to 87% as the average number of cluster users increases from 10 to 80. Although the performance of the macro cell BS_1 deteriorates as some cluster users are offloaded to BS_1, the overall mean user rate can also be improved by the proposed scheme due to more balanced traffic.

It can be further observed from Fig. 6.6(c) that the proposed scheme achieves a better overall mean user rate than the per-tier adjustment scheme. For the per-tier adjustment scheme, the overall mean user rate is even lower than that without any adjustment. Intuitively, as is illustrated in Fig. 6.3, the optimal adjustment of the per-tier BSs would offload less cluster users to the macro cell than the adjustment of BS_0, which would lead to worse performance of each user in the overloaded cell. In addition, as the association regions of all Tier-2 BSs are reduced according to the per-tier bias adjustment, the macro BS would have to undertake more traffic pressure from these lightly-loaded micro cells, and thus these cells will be more likely to become void. In this case, the BSs of the void cell will only act as interfering sources, which offsets the gain from load balancing. Compared

to the per-tier bias adjustment, the proposed scheme does not need coordinations among BSs, and is thus easy to implement in practice.

6.6 Conclusions

In this chapter we consider heterogeneous user distribution in HetNets, in particular, one cell becomes overloaded with a larger user density than other cells, and propose to adjust the biasing factor to relieve the traffic pressure of the overloaded cell. The optimal biasing factor of the overloaded cell is obtained by maximizing the mean utility of a random user in the overloaded cell and its neighboring cells. Analytical results are demonstrated by a simple case when the overloaded cell is fully surrounded by a macro cell. It is found that in this case the optimal biasing factor can be perfectly fitted as a log-linear function of the average number of cluster users. Numerical results indicate that with a proper tuning of the biasing factor of the overloaded BS, the mean user rate in the overloaded BS and the overall mean user rate can be improved. A comparative study further indicates a much better performance than the per-tier adjustment of the biasing factor.

Note that although we assume only one overloaded cell in this chapter, our proposed scheme can be extended to scenarios of multiple overloaded cells as long as these overloaded cells are not adjacent to each other and have different neighboring cells. This is because adjusting the biasing factor of each one of these overloaded cells will only affect the BS-user association in its own neighboring cells. However, for the cases where the overloaded cells appear closely and even become adjacent to each other, the adjustment of the biasing factor of one cell could have an impact on the other overloaded cells. Therefore, these related cells should cooperate with each other to obtain their optimal biasing factors, which is the next step of our study. In addition, it is assumed in this chapter that the cluster of users ideally appear in the association region of one cell. In

practice, the coverage of the cluster may overlap with several cells spatially. In this case, a more intriguing approach of a joint adjustment of the biasing factors of these related cells should be considered, and deserves much attention in the future study.

In addition, besides tuning the biasing factor of the overloaded BS, bandwidth allocation which has been studied in Chapter 4 can also be applied in this case. To be specific, when a BS is overloaded due to the cluster of user, more bandwidth can be allocated from the neighboring idle BSs to this overloaded BS, which would also involve a BS coordination issue.

Chapter 7

Conclusions and Future Works

7.1 Conclusions

The ongoing maturation of the heterogeneous network has elevated the HetNet's performance optimization to a central problem. Most of the traditional studies on the HetNet assumed continuous BS transmission, which leads to a consistent interference pattern. In addition, they characterized the locations of the users by uniform distributions. To address these open challenges, this thesis optimizes both the network spectrum efficiency and the network energy efficiency under a more practical scenario of queuing and non-uniform user distribution. The queuing behaviors of the BSs are first decoupled by adopting stochastic geometry and independent thinning approach with both spectrum partitioning and universal frequency reuse. Based on the queuing analysis, an optimal biased association scheme between users and BSs are obtained by minimizing a lower bound of the network mean queuing delay. The queue-aware optimal bandwidth allocation strategy is then studied to minimize the network average power consumption and maximize the network SIR coverage, respectively. By properly tuning the deployment density of micro BSs, the average power consumption is minimized while guaranteeing the QoS constraints

of network mean queuing delay and network SIR coverage. As last, with the consideration of non-uniform user distribution, a user intensity oriented biased scheme is proposed and studied. The contributions of this thesis is summarized as follows.

1) Queuing analysis. The queuing model in this thesis is first demonstrated and the coupled nature of the queues is examined. To solve the coupled queue problem, mathematical approaches of stochastic geometry and independent thinning are then introduced, based upon which the average traffic intensity of each tier for the cases of orthogonal spectrum partitioning and universal frequency reuse are characterized. It is shown that with spectrum partitioning, explicit expression of the average traffic intensity can be derived. With universal frequency reuse, the average traffic intensity of each tier depends on each tier, which forms a set of fixed-point equations. To justify the proceeding analysis, a spatial-temporal simulation is conducted, which indicates that the average traffic intensity of each tier can be well predicted by the adopted approaches.

2) Queue-aware delay-optimal biased association optimization in HetNets. Based on the derived expression of the average traffic intensity of each tier, the lower bound of the network mean queuing delay is characterized. The minimization problem of the lower bound of the network mean queuing delay is then formulated, which is then shown to be convex with respect to the biasing factor of each tier. When the mean packet arrival rate of each user is small, an explicit expression of the optimal biasing factor of each tier is obtained. With equal bandwidth allocation across tiers, it is further shown that each user should associate with its nearest BS. Simulation results justify our analysis by illustrating that the network mean queuing delay can be significantly reduced by a proper tuning of the biasing factor of each tier. Furthermore, by comparing the network mean queuing delay with the network SIR coverage, a tradeoff between them is revealed, indicating that a balance should be stroke between the performance of real-time and non-real-time services.

3) Queue-aware optimal bandwidth allocation in HetNets. To optimize the network energy efficiency with queuing considered, a network average power consumption minimization problem is first formulated, and is proved to be convex with respect to the bandwidth allocated to each tier. By using the approximation of the average traffic intensity of each tier, an explicit solution of the bandwidth allocated to each tier is then derived, which increases as the transmission power or the deployment density of the BSs of this tier increases. To improve the network spectrum efficiency, a maximization problem of the network SIR coverage is then studied, which is shown to be concave with respect to the bandwidth allocation. Similarly by using the approximation of the average traffic intensity, closed-form solution is obtained. It is further shown that when the mean packet arrival of each user is small, the optimal bandwidth allocated to each tier also increases as the transmission power or the deployment density of the BSs of this tier increases. Simulation results demonstrate that both the network average power consumption and SIR coverage can be remarkably improved by a proper bandwidth allocation strategy. At last, a tradeoff is revealed between the network energy efficiency and SIR coverage.

4) Queue-aware energy efficient BS density optimization in HetNets. In the consideration of universal frequency reuse by a 2-Tier HetNet, the existence and uniqueness of the fixed-point equations of the average traffic intensity of each tier is proved. To numerically obtain the solution, an iterative method is proposed and its convergence is then proved. By further using the approximation that BSs of a tier have the same SIR coverage, the CDF of the traffic intensity of each tier is obtained. On that basis, a network average power consumption minimization problem under the constraints of the network mean queuing delay and the network SIR coverage is formulated. Numerical results show that if the idle power coefficient is below a certain threshold, the optimal activation ratio should equal the one to minimize the network average power consumption per area. Otherwise, the optimal activation ratio should be obtained according to the QoS constraints. It is fur-

ther revealed that by taking queuing into account, universal spectrum reuse outperforms spectrum partitioning in terms of energy efficiency and SIR coverage in the considered scenario.

5) Optimal biased association scheme with non-uniform user distribution. The scenario that one cell is overloaded with a cluster of users, i.e., has a higher user intensity, is examined. A load-aware biased association scheme where the overloaded BS optimally adjusts its biasing factor according to its load condition by maximizing a utility function of the mean user rate. By studying the case where one micro BS is fully surrounded by a macro BS, we find that the optimal biasing factor of the overloaded BS can be perfectly fitted as a log-linear function of its user density of the overloaded cell, which greatly facilitates the implementation of our proposed scheme. Simulation results demonstrate the proposed scheme can significantly improve the mean user rate in the overloaded cell, and the overall mean user rate performance can also be improved due to a more balanced load.

7.2 Future Works

This thesis addresses the performance optimization of HetNets by considering queuing dynamics and a more practical user distribution model. The achieved results in this thesis can lead to many interesting open questions. Some directions for future research are pointed out as follows.

1) Combination of queuing dynamics and non-uniform user distribution. Although this thesis considers the more practical scenario of queuing dynamics and non-uniform user distribution, these two assumptions are not combined together, which could result in a different and more complex coupled queue problem. In particular, due to the cluster of users, the traffic intensity of the overloaded cell should be much more higher than that of

the regular cell in one tier. Therefore, the neighboring BSs of such overloaded BS have a higher level of the experienced interference, which in turns raises the busy probability of the overloaded BS. The queuing analysis of the overloaded BS as well as its neighboring BSs remains an interesting issue that attracts much attention in the future.

2) Uplink transmission model. Quite different from the downlink transmission where the interfering sources are BSs with high transmission powers, the source of the interference in the uplink are the mobile users with low levels of power and more diverse geographical distribution. Although the homogeneous PPP assumption for BSs of each tier greatly simplifies the downlink interference characterization, analysis of the uplink in such a setting is highly non-trivial, as the uplink interference does not originate from Poisson distributed nodes. Therefore, a mathematical model needs to be adopted to characterize the locations of the mobile users. Moreover, as power control is usually adopted in the uplink to fully or partially compensate for the path loss, users may choose to associate with different BSs in the uplink. Therefore, there exists a decoupled association strategy for the users between downlink and uplink, and such user-BS association optimization becomes more intriguing in the uplink. At last, as each mobile user now has its own queue in the uplink, characterizing the traffic intensity of each individual user is thus more complicated. Hence, how to optimize the network performance for the uplink in terms of both network spectrum efficiency and energy efficiency deserves much attention in the future study.

Appendix A

Abbreviations

5G	Fifth Generation
AP	Access Point
BS	Base Station
CDF	Cumulative Distribution Function
CDMA	Code Division Multiple Access
CRE	Cell Range Expansion
CSI	Channel State Information
CTMT	Continuous Time Markov Chain
FIFS	First In First Serve
FUA	Fractional User Association
HetNet	Heterogeneous Network
LTE	Long Term Evolution
mmWave	millimeter wave
MDP	Markov Decision Problem
MIMO	Multiple Input Multiple Output
PDF	Probability Density Function

PGFL	Probability Generating Functional
PPP	Poisson Point Process
QoS	Quality of Service
RAT	Radio Access Technology
RSRP	Reference Signal Receiving Power
RSSI	Received Signal Strength Indicator
SIR	Signal to Interference Ratio
SINR	Signal to Interference plus Noise Ratio
SP	Spectrum Partitioning
UFR	Universal Frequency Reuse
VNI	Visual Network Index
WLAN	Wireless Local Area Network

Appendix B

Publications

Some of the work presented in this thesis has appeared as published journal and conference papers, has been accepted for publication at the time of writing this thesis, or has been submitted for publication at the time of writing this thesis. A list of the papers is provided below.

[**P1**] Fancheng Kong, Xinghua Sun, Hongbo Zhu, "Optimal biased association scheme with heterogeneous user distribution in HetNets," *Wireless Personal Communications*, vol. 90, no. 2, pp. 575–594, Sept. 2016.

[**P2**] Fancheng Kong, Xinghua Sun, Victor C. M. Lueng, Hongbo Zhu, "Delay-optimal biased user association in heterogeneous network," *IEEE Transactions on Vehicular Technology*, vol. 66, no. 8, pp. 7360–7371, Feb. 2017.

[**P3**] Fancheng Kong, Xinghua Sun, Victor. C. M. Leung, Yingjie. J. Guo, Qi Zhu, Hongbo Zhu, "Queue-aware small cell activation for energy efficiency in two-tier heterogeneous networks," *2017 IEEE Wireless Communications and Networking Conference (WCNC)*, San Francisco, CA, 2017, pp. 1-6.

[**P4**] Fancheng Kong, Xinghua Sun, Yingjie. J. Guo, Hongbo Zhu, "Queue-aware optimal

bandwidth allocation in heterogeneous networks," *IEEE Wireless Communication Letters*, vol. 6, no. 6, pp. 730–733, Aug. 2017.

[**P5**] Fancheng Kong, Xinghua Sun, Victor. C. M. Leung, Yingjie. J. Guo, Qi Zhu, Hongbo Zhu, "Queue-aware power consumption minimization in two-tier heterogeneous networks," *To appear in IEEE Transactions on Vehicular Technology.*

Bibliography

[1] J. G. Andrews, S. Buzzi, W. Choi, S. V. Hanly, A. Lozano, A. C. K. Soong, and J. C. Zhang, "What will 5G be?" *IEEE Journal on Selected Areas in Communications*, vol. 32, no. 6, pp. 1065–1082, June 2014.

[2] Cisco, "Global mobile data traffic forecast update, 2016–2021," [Online]. Available: https://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/mobile-white-paper-c11-520862.

[3] Qualcomm, "The 1000x mobile data challenge," 2015, [Online]. Available: https://www.qualcomm.com/invention/1000x.

[4] A. Maeder, P. Rost, and D. Staehle, "The challenge of M2M communications for the cellular radio access network," in *Proceedings of Wurzburg Workshop IP*. Joint ITG Euro-NF Workshop, 2011.

[5] M. S. Corson, R. Laroia, J. Li, V. Park, T. Richardson, and G. Tsirtsis, "Toward proximity-aware internetworking," *IEEE Wireless Communications*, vol. 17, no. 6, pp. 26–33, Dec. 2010.

[6] S. Hilton, "Machine-to-machine device connections: Worldwide forecast 2010–2020," 2010, Analysys Mason Report.

[7] FP7 European Project 317669 METIS (Mobile and Wireless Communications Enablers for the Twenty-Twenty Information Society) 2012, [Online]. Available: https://www.metis2020.com.

[8] T. L. Marzetta, "Noncooperative cellular wireless with unlimited numbers of base station antennas," *IEEE Transactions on Wireless Communications*, vol. 9, no. 11, pp. 3590–3600, November 2010.

[9] J. Hoydis, S. ten Brink, and M. Debbah, "Massive MIMO in the UL/DL of cellular networks: How many antennas do we need?" *IEEE Journal on Selected Areas in Communications*, vol. 31, no. 2, pp. 160–171, February 2013.

[10] F. Rusek, D. Persson, B. K. Lau, E. G. Larsson, T. L. Marzetta, O. Edfors, and F. Tufvesson, "Scaling up MIMO: Opportunities and challenges with very large arrays," *IEEE Signal Processing Magazine*, vol. 30, no. 1, pp. 40–60, Jan 2013.

[11] D. C. Araujo, T. Maksymyuk, A. L. F. de Almeida, T. Maciel, J. C. M. Mota, and M. Jo, "Massive MIMO: survey and future research topics," *IET Communications*, vol. 10, no. 15, pp. 1938–1946, 2016.

[12] E. G. Larsson, O. Edfors, F. Tufvesson, and T. L. Marzetta, "Massive MIMO for next generation wireless systems," *IEEE Communications Magazine*, vol. 52, no. 2, pp. 186–195, February 2014.

[13] L. Lu, G. Y. Li, A. L. Swindlehurst, A. Ashikhmin, and R. Zhang, "An overview of massive MIMO: Benefits and challenges," *IEEE Journal of Selected Topics in Signal Processing*, vol. 8, no. 5, pp. 742–758, Oct 2014.

[14] M. Marcus and B. Pattan, "Millimeter wave propagation; spectrum management implications," *IEEE Microwave Magazine*, vol. 6, no. 2, pp. 54–62, June 2005.

[15] A. V. Alejos, M. G. Sanchez, and I. Cuinas, "Measurement and analysis of propagation mechanisms at 40 GHz: Viability of site shielding forced by obstacles," *IEEE Transactions on Vehicular Technology*, vol. 57, no. 6, pp. 3369–3380, Nov 2008.

[16] Z. Pi and F. Khan, "An introduction to millimeter-wave mobile broadband systems," *IEEE Communications Magazine*, vol. 49, no. 6, pp. 101–107, June 2011.

[17] T. S. Rappaport, S. Sun, R. Mayzus, H. Zhao, Y. Azar, K. Wang, G. N. Wong, J. K. Schulz, M. Samimi, and F. Gutierrez, "Millimeter wave mobile communications for 5G cellular: It will work!" *IEEE Access*, vol. 1, pp. 335–349, 2013.

[18] W. Roh, J. Y. Seol, J. Park, B. Lee, J. Lee, Y. Kim, J. Cho, K. Cheun, and F. Aryanfar, "Millimeter-wave beamforming as an enabling technology for 5G cellular communications: theoretical feasibility and prototype results," *IEEE Communications Magazine*, vol. 52, no. 2, pp. 106–113, February 2014.

[19] S. Rangan, T. S. Rappaport, and E. Erkip, "Millimeter-wave cellular wireless networks: Potentials and challenges," *Proceedings of the IEEE*, vol. 102, no. 3, pp. 366–385, March 2014.

[20] J. G. Andrews, H. Claussen, M. Dohler, S. Rangan, and M. C. Reed, "Femtocells: Past, present, and future," *IEEE Journal on Selected Areas in Communications*, vol. 30, no. 3, pp. 497–508, April 2012.

[21] V. Chandrasekhar, J. G. Andrews, and A. Gatherer, "Femtocell networks: a survey," *IEEE Communications Magazine*, vol. 46, no. 9, pp. 59–67, September 2008.

[22] H. Claussen, L. T. W. Ho, and L. G. Samuel, "An overview of the femtocell concept," *Bell Labs Technical Journal*, vol. 13, no. 1, pp. 221–245, Spring 2008.

[23] I. G. T. Q. S. Quek, G. de la Roche and M. Kountouris, *Small Cell Networks: Deployment, PHY Techniques, and Resource Management.* Cambridge University Press, 2013.

[24] N. Bhushan, J. Li, D. Malladi, R. Gilmore, D. Brenner, A. Damnjanovic, R. T. Sukhavasi, C. Patel, and S. Geirhofer, "Network densification: the dominant theme for wireless evolution into 5G," *IEEE Communications Magazine*, vol. 52, no. 2, pp. 82–89, February 2014.

[25] D. P. Malladi, "Heterogeneous networks in 3G and 4G," in *Proceedings of IEEE Communication Theory Workshop*, May 2012, http://www.ieeectw.org/program.

[26] X. Lagrange, "Multitier cell design," *IEEE Communications Magazine*, vol. 35, no. 8, pp. 60–64, Aug 1997.

[27] A. A. M. Saleh, A. Rustako, and R. Roman, "Distributed antennas for indoor radio communications," *IEEE Transactions on Communications*, vol. 35, no. 12, pp. 1245–1251, December 1987.

[28] J. Sydir and R. Taori, "An evolved cellular system architecture incorporating relay stations," *IEEE Communications Magazine*, vol. 47, no. 6, pp. 115–121, June 2009.

[29] X. Wu, B. Murherjee, and D. Ghosal, "Hierarchical architectures in the third-generation cellular network," *IEEE Wireless Communications*, vol. 11, no. 3, pp. 62–71, June 2004.

[30] J. G. Andrews, S. Singh, Q. Ye, X. Lin, and H. S. Dhillon, "An overview of load balancing in hetnets: old myths and open problems," *IEEE Wireless Communications*, vol. 21, no. 2, pp. 18–25, April 2014.

[31] J. G. Andrews, "Seven ways that HetNets are: a cellular paradigm shift," *IEEE Communications Magazine*, vol. 51, no. 3, pp. 136–144, March 2013.

[32] Qualcomm, "A comparison of LTE-Advanced hetnets and wifi," white paper, 2011, [Online]. Available: https://goo.gl/BFMFR.

[33] Juniper, "Wifi and femtocell integration strategies 2011–2015," white paper, 2011, [Online]. Available: https://www.juniperresearch.com/.

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

[35] S. Singh, J. G. Andrews, and G. de Veciana, "Interference shaping for improved quality of experience for real-time video streaming," *IEEE Journal on Selected Areas in Communications*, vol. 30, no. 7, pp. 1259–1269, August 2012.

[36] C. H. Liu and L. C. Wang, "Random cell association and void probability in poisson-distributed cellular networks," in *Proceedings of IEEE International Conference on Communications (ICC)*, June 2015, pp. 2816–2821.

[37] C. T. Peng, L. C. Wang, and C. H. Liu, "Optimal base station deployment for small cell networks with energy-efficient power control," in *Proceedings of IEEE International Conference on Communications (ICC)*, June 2015, pp. 1863–1868.

[38] M. Vajapeyam, A. Damnjanovic, J. Montojo, T. Ji, Y. Wei, and D. Malladi, "Downlink FTP performance of heterogeneous networks for LTE-Advanced," in *Proceedings of IEEE International Conference on Communications Workshops (ICC)*, June 2011, pp. 1–5.

[39] Y. Wang and K. I. Pedersen, "Performance analysis of enhanced inter-cell interference coordination in LTE-Advanced heterogeneous networks," in *Proceedings of IEEE 75th Vehicular Technology Conference (VTC)*, May 2012, pp. 1–5.

[40] A. Barbieri, P. Gaal, S. Geirhofer, T. Ji, D. Malladi, Y. Wei, and F. Xue, "Coordinated downlink multi-point communications in heterogeneous cellular networks," in *Proceedings of Information Theory and Applications Workshop*, Feb 2012, pp. 7–16.

[41] Nokia, "Aspects of pico node range extension," Nokia Siemens Networks, 3GPP TSG RAN WG1 meeting 61, R1–103824, 2010, [Online]. Available: http://goo.gl/XDKXI.

[42] Y. Wang, B. Soret, and K. I. Pedersen, "Sensitivity study of optimal eICIC configurations in different heterogeneous network scenarios," in *Proceedings of IEEE International Conference on Communications (ICC)*, June 2012, pp. 6792–6796.

[43] T. Nihtila and V. Haikola, "Hsdpa performance with dual stream MIMO in a combined macro-femto cell network," in *Proceedings of IEEE 71st Vehicular Technology Conference (VTC)*, May 2010, pp. 1–5.

[44] H. R. Karimi, L. T. W. Ho, H. Claussen, and L. G. Samuel, "Evolution towards dynamic spectrum sharing in mobile communications," in *Proceedings of IEEE 17th International Symposium on Personal, Indoor and Mobile Radio Communications*, Sept 2006, pp. 1–5.

[45] Qualcomm, "A 3G/LTE Wi-Fi offload framework," white paper, 2011, [Online]. Available: http://goo.gl/91EqQ.

[46] Q. Ye, B. Rong, Y. Chen, M. Al-Shalash, C. Caramanis, and J. G. Andrews, "User association for load balancing in heterogeneous cellular networks," *IEEE Transactions on Wireless Communications*, vol. 12, no. 6, pp. 2706–2716, June 2013.

[47] A. Bedekar and R. Agrawal, "Optimal muting and load balancing for eICIC," in *Proceedings of 11th International Symposium and Workshops on Modeling and Optimization in Mobile, Ad Hoc and Wireless Networks (WiOpt)*, May 2013, pp. 280–287.

[48] Q. Ye, M. Al-Shalash, C. Caramanis, and J. G. Andrews, "On/off macrocells and load balancing in heterogeneous cellular networks," in *Proceedings of IEEE Global Communications Conference (GLOBECOM)*, Dec 2013, pp. 3814–3819.

[49] S. Deb, P. Monogioudis, J. Miernik, and J. P. Seymour, "Algorithms for enhanced inter-cell interference coordination (eICIC) in LTE hetnets," *IEEE/ACM Transactions on Networking*, vol. 22, no. 1, pp. 137–150, Feb 2014.

[50] W. Li, S. Wang, Y. Cui, X. Cheng, R. Xin, M. A. Al-Rodhaan, and A. Al-Dhelaan, "AP association for proportional fairness in multirate WLANs," *IEEE/ACM Transactions on Networking*, vol. 22, no. 1, pp. 191–202, Feb 2014.

[51] L. Chen and H. Li, "An MDP-based vertical handoff decision algorithm for heterogeneous wireless networks," in *Proceedings of IEEE Wireless Communications and Networking Conference (WCNC)*, April 2016, pp. 1–6.

[52] S. E. Elayoubi, E. Altman, M. Haddad, and Z. Altman, "A hybrid decision approach for the association problem in heterogeneous networks," in *Proceedings of IEEE International Conference on Computer Communications (INFOCOM)*, March 2010, pp. 1–5.

[53] J. Zhu, L. Xu, L. Yang, and W. Xie, "An optimal vertical handoff decision algorithm for multiple services with different priorities in heterogeneous wireless networks," *Wireless Personal Communications*, vol. 83, no. 1, pp. 527–549, Jul 2015. [Online]. Available: https://doi.org/10.1007/s11277-015-2407-1

[54] E. Aryafar, A. Keshavarz-Haddad, M. Wang, and M. Chiang, "RAT selection games in hetnets," in *Proceedings of IEEE International Conference on Computer Communications (INFOCOM)*, April 2013, pp. 998–1006.

[55] D. Niyato and E. Hossain, "Dynamics of network selection in heterogeneous wireless networks: An evolutionary game approach," *IEEE Transactions on Vehicular Technology*, vol. 58, no. 4, pp. 2008–2017, May 2009.

[56] S. Das, H. Viswanathan, and G. Rittenhouse, "Dynamic load balancing through coordinated scheduling in packet data systems," in *Proceedings of IEEE International Conference on Computer Communications (INFOCOM)*, vol. 1, March 2003, pp. 786–796 vol.1.

[57] I. UlHaq, A. Shafiq, K. M. Yahya, and M. N. Iqbal, "Cell breathing and cell capacity in CDMA: algorithm amp; evaluation," in *Proceedings of 7th International Symposium on Communication Systems, Networks Digital Signal Processing (CSNDSP)*, July 2010, pp. 432–436.

[58] A. Sang, X. Wang, M. Madihian, and R. D. Gitlin, "Coordinated load balancing, handoff/cell-site selection, and scheduling in multi-cell packet data systems," *Wireless Networks*, vol. 14, no. 1, pp. 103–120, Feb 2008. [Online]. Available: https://doi.org/10.1007/s11276-006-8533-7

[59] H. S. Jo, Y. J. Sang, P. Xia, and J. G. Andrews, "Heterogeneous cellular networks with flexible cell association: A comprehensive downlink SINR analysis," *IEEE Transactions on Wireless Communications*, vol. 11, no. 10, pp. 3484–3495, October 2012.

[60] S. Mukherjee and I. Guvenc, "Effects of range expansion and interference coordination on capacity and fairness in heterogeneous networks," in *Proceedings of*

Conference Record of the Forty Fifth Asilomar Conference on Signals, Systems and Computers (ASILOMAR), Nov 2011, pp. 1855–1859.

[61] D. Lopez-Perez, X. Chu, and I. Guvenc, "On the expanded region of picocells in heterogeneous networks," *IEEE Journal of Selected Topics in Signal Processing*, vol. 6, no. 3, pp. 281–294, June 2012.

[62] T. Bu, L. Li, and R. Ramjee, "Generalized proportional fair scheduling in third generation wireless data networks," in *Proceedings of 25th IEEE International Conference on Computer Communications (INFOCOM)*, April 2006, pp. 1–12.

[63] M. Haenggi, *Stochastic Geometry for Wireless Networks*. Cambridge University Press, 2012.

[64] M. Haenggi, J. G. Andrews, F. Baccelli, O. Dousse, and M. Franceschetti, "Stochastic geometry and random graphs for the analysis and design of wireless networks," *IEEE Journal on Selected Areas in Communications*, vol. 27, no. 7, pp. 1029–1046, September 2009.

[65] H. S. Dhillon, R. K. Ganti, F. Baccelli, and J. G. Andrews, "Modeling and analysis of K-Tier downlink heterogeneous cellular networks," *IEEE Journal on Selected Areas in Communications*, vol. 30, no. 3, pp. 550–560, April 2012.

[66] S. Singh, H. S. Dhillon, and J. G. Andrews, "Offloading in heterogeneous networks: Modeling, analysis, and design insights," *IEEE Transactions on Wireless Communications*, vol. 12, no. 5, pp. 2484–2497, May 2013.

[67] S. Singh and J. G. Andrews, "Joint resource partitioning and offloading in heterogeneous cellular networks," *IEEE Transactions on Wireless Communications*, vol. 13, no. 2, pp. 888–901, February 2014.

[68] Y. Lin, W. Bao, W. Yu, and B. Liang, "Optimizing user association and spectrum allocation in hetnets: A utility perspective," *IEEE Journal on Selected Areas in Communications*, vol. 33, no. 6, pp. 1025–1039, June 2015.

[69] M. Marchese, *QoS Over Heterogeneous Networks*. Wiley Publishing, 2007.

[70] P. Y. Kong, "Power consumption and packet delay relationship for heterogeneous wireless networks," *IEEE Communications Letters*, vol. 17, no. 7, pp. 1376–1379, July 2013.

[71] M. Simsek, A. Aijaz, M. Dohler, J. Sachs, and G. Fettweis, "5G-enabled tactile internet," *IEEE Journal on Selected Areas in Communications*, vol. 34, no. 3, pp. 460–473, March 2016.

[72] G. P. Fettweis, "The tactile internet: Applications and challenges," *IEEE Vehicular Technology Magazine*, vol. 9, no. 1, pp. 64–70, March 2014.

[73] "End-to-end quality of service (QoS) concept and architecture," 2012, 3GPP TS 23.107 version 11.0.0 Release 11.

[74] R. Baldemair, E. Dahlman, G. Fodor, G. Mildh, S. Parkvall, Y. Selen, H. Tullberg, and K. Balachandran, "Evolving wireless communications: Addressing the challenges and expectations of the future," *IEEE Vehicular Technology Magazine*, vol. 8, no. 1, pp. 24–30, March 2013.

[75] C. X. Wang, F. Haider, X. Gao, X. H. You, Y. Yang, D. Yuan, H. M. Aggoune, H. Haas, S. Fletcher, and E. Hepsaydir, "Cellular architecture and key technologies for 5G wireless communication networks," *IEEE Communications Magazine*, vol. 52, no. 2, pp. 122–130, February 2014.

[76] S. Shioda, "Fundamental trade-offs between resource separation and resource share for quality of service guarantees," *IET Networks*, vol. 3, no. 1, pp. 4–15, March 2014.

[77] S. M. Saheb, A. K. Bhattacharjee, P. Dharmasa, and R. Kar, "Enhanced hybrid coordination function controlled channel access-based adaptive scheduler for delay sensitive traffic in IEEE 802.11e networks," *IET Networks*, vol. 1, no. 4, pp. 281–288, December 2012.

[78] M. Haenggi, "The local delay in poisson networks," *IEEE Transactions on Information Theory*, vol. 59, no. 3, pp. 1788–1802, March 2013.

[79] Z. Gong and M. Haenggi, "The local delay in mobile poisson networks," *IEEE Transactions on Wireless Communications*, vol. 12, no. 9, pp. 4766–4777, September 2013.

[80] T. Bonald and A. Proutire, "Wireless downlink data channels: user performance and cell dimensioning," in *Proceedings of International Conference on Mobile Computing and Networking*, 2003, pp. 339–352.

[81] S. Borst, "User-level performance of channel-aware scheduling algorithms in wireless data networks," *IEEE/ACM Transactions on Networking*, vol. 13, no. 3, pp. 636–647, June 2005.

[82] D. C. Chen, T. Q. S. Quek, and M. Kountouris, "Backhauling in heterogeneous cellular networks: Modeling and tradeoffs," *IEEE Transactions on Wireless Communications*, vol. 14, no. 6, pp. 3194–3206, June 2015.

[83] B. Zhuang, D. Guo, and M. L. Honig, "Traffic driven resource allocation in heterogenous wireless networks," in *Proceedings of IEEE Global Communications Conference (GLOBECOM)*, Dec 2014, pp. 1546–1551.

[84] ——, "Traffic-driven spectrum allocation in heterogeneous networks," *IEEE Journal on Selected Areas in Communications*, vol. 33, no. 10, pp. 2027–2038, Oct 2015.

[85] A. Cheng, J. Li, Y. Yu, and H. Jin, "Delay-sensitive user scheduling and power control in heterogeneous networks," *IET Networks*, vol. 4, no. 3, pp. 175–184, 2015.

[86] B. Rengarajan and G. de Veciana, "Architecture and abstractions for environment and traffic-aware system-level coordination of wireless networks," *IEEE/ACM Transactions on Networking*, vol. 19, no. 3, pp. 721–734, June 2011.

[87] S. Borst, N. Hegde, and A. Proutire, "Interacting queues with server selection and coordinated schedulingapplication to cellular data networks," *Annals of Operations Research*, vol. 170, no. 1, pp. 59–78, 2009.

[88] H. S. Dhillon, R. K. Ganti, and J. G. Andrews, "Load-aware modeling and analysis of heterogeneous cellular networks," *IEEE Transactions on Wireless Communications*, vol. 12, no. 4, pp. 1666–1677, April 2013.

[89] I. J. B. F. Adan, "A compensation approach for two-dimensional Markov processes," *Advances in Applied Probability*, vol. 25, no. 4, pp. 783–817, 1993.

[90] A. Fehske, G. Fettweis, J. Malmodin, and G. Biczok, "The global footprint of mobile communications: The ecological and economic perspective," *IEEE Communications Magazine*, vol. 49, no. 8, pp. 55–62, August 2011.

[91] L. M. Correia, D. Zeller, O. Blume, D. Ferling, Y. Jading, I. Gdor, G. Auer, and L. V. D. Perre, "Challenges and enabling technologies for energy aware mobile radio networks," *IEEE Communications Magazine*, vol. 48, no. 11, pp. 66–72, November 2010.

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

[93] E. Oh, B. Krishnamachari, X. Liu, and Z. Niu, "Toward dynamic energy-efficient operation of cellular network infrastructure," *IEEE Communications Magazine*, vol. 49, no. 6, pp. 56–61, June 2011.

[94] S. Mclaughlin, P. M. Grant, J. S. Thompson, H. Haas, D. I. Laurenson, C. Khirallah, Y. Hou, and R. Wang, "Techniques for improving cellular radio base station energy efficiency," *IEEE Wireless Communications*, vol. 18, no. 5, pp. 10–17, October 2011.

[95] G. Auer, V. Giannini, C. Desset, I. Godor, P. Skillermark, M. Olsson, M. A. Imran, D. Sabella, M. J. Gonzalez, O. Blume, and A. Fehske, "How much energy is needed to run a wireless network?" *IEEE Wireless Communications*, vol. 18, no. 5, pp. 40–49, October 2011.

[96] R. Q. Hu and Y. Qian, "An energy efficient and spectrum efficient wireless heterogeneous network framework for 5G systems," *IEEE Communications Magazine*, vol. 52, no. 5, pp. 94–101, May 2014.

[97] L. An, T. Zhang, and C. Feng, "Joint optimization for base station density and user association in energy-efficient cellular networks," in *Proceedings of International Symposium on Wireless Personal Multimedia Communications (WPMC)*, Sept 2014, pp. 85–90.

[98] ——, "Stochastic geometry based energy-efficient base station density optimization in cellular networks," in *Proceedings of IEEE Wireless Communications and Networking Conference (WCNC)*, March 2015, pp. 1614–1619.

[99] S. Sarkar, R. K. Ganti, and M. Haenggi, "Optimal base station density for power efficiency in cellular networks," in *Proceedings of IEEE International Conference on Communications (ICC)*, June 2014, pp. 4054–4059.

[100] E. Oh, K. Son, and B. Krishnamachari, "Dynamic base station switching-on/off strategies for green cellular networks," *IEEE Transactions on Wireless Communications*, vol. 12, no. 5, pp. 2126–2136, May 2013.

[101] W. Guo and T. O'Farrell, "Dynamic cell expansion: Traffic aware low energy cellular network," in *Proceedings of IEEE Vehicular Technology Conference (VTC)*, Sept 2012, pp. 1–5.

[102] J. Wu, Y. Zhang, M. Zukerman, and E. K. N. Yung, "Energy-efficient base-stations sleep-mode techniques in green cellular networks: A survey," *IEEE Communications Surveys Tutorials*, vol. 17, no. 2, pp. 803–826, Secondquarter 2015.

[103] S. Cai, Y. Che, L. Duan, J. Wang, S. Zhou, and R. Zhang, "Green 5G heterogeneous networks through dynamic small-cell operation," *IEEE Journal on Selected Areas in Communications*, vol. 34, no. 5, pp. 1103–1115, May 2016.

[104] B. Zhuang, D. Guo, and M. L. Honig, "Energy-efficient cell activation, user association, and spectrum allocation in heterogeneous networks," *IEEE Journal on Selected Areas in Communications*, vol. 34, no. 4, pp. 823–831, April 2016.

[105] L. Li, M. Peng, C. Yang, Y. Wu, W. Xue, and Y. Li, "Base station density optimization for high energy efficiency in two-tier cellular networks," in *Proceedings of IEEE Global Communications Conference (GLOBECOM)*, Dec 2014, pp. 1804–1809.

[106] L. Li, M. Peng, C. Yang, and Y. Wu, "Optimization of base-station density for high energy-efficient cellular networks with sleeping strategies," *IEEE Transactions on Vehicular Technology*, vol. 65, no. 9, pp. 7501–7514, Sept 2016.

[107] D. Cao, S. Zhou, and Z. Niu, "Optimal combination of base station densities for energy-efficient two-tier heterogeneous cellular networks," *IEEE Transactions on Wireless Communications*, vol. 12, no. 9, pp. 4350–4362, September 2013.

[108] ——, "Improving the energy efficiency of two-tier heterogeneous cellular networks through partial spectrum reuse," *IEEE Transactions on Wireless Communications*, vol. 12, no. 8, pp. 4129–4141, August 2013.

[109] P. Frenger, P. Moberg, J. Malmodin, Y. Jading, and I. Godor, "Reducing energy consumption in LTE with cell DTX," in *Proceedings of IEEE 73rd Vehicular Technology Conference (VTC)*, May 2011, pp. 1–5.

[110] J. G. Andrews, F. Baccelli, and R. K. Ganti, "A tractable approach to coverage and rate in cellular networks," *IEEE Transactions on Communications*, vol. 59, no. 11, pp. 3122–3134, November 2011.

[111] B. Blaszczyszyn, M. K. Karray, and H. P. Keeler, "Using poisson processes to model lattice cellular networks," in *Proceedings of IEEE International Conference on Computer Communications (INFOCOM)*, April 2013, pp. 773–781.

[112] D. Stoyan, W. S. Kendall, and J. Mecke, *Stochastic Geometry and Its Applications*. John Wiley & Sons, 1995.

[113] C. H. M. de Lima, M. Bennis, and M. Latva-aho, "Statistical analysis of self-organizing networks with biased cell association and interference avoidance," *IEEE Transactions on Vehicular Technology*, vol. 62, no. 5, pp. 1950–1961, Jun 2013.

[114] F. Kong, X. Sun, and H. Zhu, "Optimal biased association scheme with heterogeneous user distribution in HetNets," *Wireless Personal Communications*, vol. 90, no. 2, pp. 575–594, Sep 2016. [Online]. Available: https://doi.org/10.1007/s11277-015-3102-y

[115] B. Badic, T. O'Farrrell, P. Loskot, and J. He, "Energy efficient radio access architectures for green radio: Large versus small cell size deployment," in *Proceedings of IEEE 70th Vehicular Technology Conference Fall (VTC)*, Sept 2009, pp. 1–5.

[116] M. A. Marsan, L. Chiaraviglio, D. Ciullo, and M. Meo, "Optimal energy savings in cellular access networks," in *Proceedings of IEEE International Conference on Communications Workshops*, June 2009, pp. 1–5.

[117] H. S. Jung, H. T. Roh, and J. W. Lee, "Energy and traffic aware dynamic topology management for wireless cellular networks," in *Proceedings of IEEE International Conference on Communication Systems (ICCS)*, Nov 2012, pp. 205–209.

[118] R. Ramamonjison, G. K. Iran, K. Sakaguchi, K. Araki, S. Kaneko, Y. Kishi, and N. Miyazaki, "Spectrum allocation strategies for heterogeneous networks," in *Proceedings of 6th International ICST Conference on Cognitive Radio Oriented Wireless Networks and Communications (CROWNCOM)*, June 2011, pp. 201–205.

[119] Y. Yan, Y. Li, C. Xing, and J. Wang, "Spectrum allocation in two-tier heterogeneous network: A bilateral negotiation framework," in *Proceedings of IEEE 81st Vehicular Technology Conference (VTC)*, May 2015, pp. 1–5.

[120] S. Sadr and R. S. Adve, "Tier association probability and spectrum partitioning for maximum rate coverage in multi-tier heterogeneous networks," *IEEE Communications Letters*, vol. 18, no. 10, pp. 1791–1794, Oct 2014.

[121] W. Bao and B. Liang, "Structured spectrum allocation and user association in heterogeneous cellular networks," in *Proceedings of IEEE Conference on Computer Communications (INFOCOM)*, April 2014, pp. 1069–1077.

[122] ——, "Rate maximization through structured spectrum allocation and user association in heterogeneous cellular networks," *IEEE Transactions on Communications*, vol. 63, no. 11, pp. 4510–4524, Nov 2015.

[123] M. Feng, S. Mao, and T. Jiang, "BOOST: Base station on-off switching strategy for energy efficient massive MIMO hetnets," in *Proceedings of IEEE International Conference on Computer Communications (INFOCOM)*, April 2016, pp. 1–9.

[124] G. Lim, C. Xiong, L. J. Cimini, and G. Y. Li, "Energy-efficient resource allocation for OFDMA-based multi-RAT networks," *IEEE Transactions on Wireless Communications*, vol. 13, no. 5, pp. 2696–2705, May 2014.

[125] S. Kim, B. G. Lee, and D. Park, "Energy-per-bit minimized radio resource allocation in heterogeneous networks," *IEEE Transactions on Wireless Communications*, vol. 13, no. 4, pp. 1862–1873, April 2014.

[126] Q. D. Vu, L. N. Tran, M. Juntti, and E. K. Hong, "Energy-efficient bandwidth and power allocation for multi-homing networks," *IEEE Transactions on Signal Processing*, vol. 63, no. 7, pp. 1684–1699, April 2015.

[127] Y. Zhong, T. Q. S. Quek, and X. Ge, "Heterogeneous cellular networks with spatio-temporal traffic: Delay analysis and scheduling," *IEEE Journal on Selected Areas in Communications*, vol. 35, no. 6, pp. 1373–1386, June 2017.

[128] G. Zhang, T. Q. S. Quek, A. Huang, and H. Shan, "Delay and reliability tradeoffs in heterogeneous cellular networks," *IEEE Transactions on Wireless Communications*, vol. 15, no. 2, pp. 1101–1113, Feb 2016.

[129] X. Luo, "Delay-oriented QoS-aware user association and resource allocation in heterogeneous cellular networks," *IEEE Transactions on Wireless Communications*, vol. 16, no. 3, pp. 1809–1822, March 2017.

[130] J. S. Ferenc and Z. Neda, "On the size distribution of poisson voronoi cells," *Physica A Statistical Mechanics & Its Applications*, vol. 385, no. 2, pp. 518–526, 2007.