Energy-efficient Strategies with Base Station Power Management for Green Wireless Networks

by

Hong Zhang

A Thesis submitted to the Faculty of Graduate Studies of
The University of Manitoba
in partial fulfillment of the requirements of the degree of

DOCTOR OF PHILOSOPHY

Department of Electrical and Computer Engineering
University of Manitoba
Winnipeg

Supervisor: **Prof. Jun Cai**

Abstract

In this thesis, our objective is to improve the energy efficiency and load balance for wireless networks. We first study the relationships between the base station (BS) on/off operation and traffic distribution. A cooperative power saving method called clustering BS-off (CBSO) scheme is proposed. Instead of adopting a unified and consistent BS-off scheme in the whole network, the proposed centralized and distributed CBSO schemes can adaptively group BSs in several clusters based on the traffic fluctuations with space and time. Second, to further improve the network load balance and energy efficiency in distributed manner, we propose a power efficient self-organized virtual small networking (VSN) protocol. A heuristic firefly algorithm is applied to arrange the BSs' operation in small groups based on the traffic level. By jointly considering the load balance, the effectiveness of the proposed algorithm is demonstrated based on the average and min-max traffic levels of BSs' groups. Finally, the importance of detailed BS operation between active and sleep modes is considered. The operating procedure of femtocell base station, i.e., HeNB, is modeled as an MAP/PH/1/k queueing system. Such queueing analysis particularly focuses on the HeNB vacation process with user priorities. The HeNB's power on/off scheme is modeled as alternative service and vacation periods. The hybrid access is regarded as high and low priority users in the queuing system. We further propose the adaptive service rate and vacation length (ASV) method, so that the HeNB can work in a more energy-efficient way while satisfying QoS requirements such as blocking probability and users waiting time. Simulation results show the effectiveness of the proposed strategies and the overall network energy efficiency can be improved significantly.

Keywords: Wireless networks, energy efficiency, base station power control, load balance, small cell networks, femtocells, hybrid access, vacation queue, priority queue.

Acknowledgments

First and foremost, I would like to express my sincere gratitude to my advisor Prof. Jun Cai for the continuous support of my Ph.D. study and related research, for his patience, motivation, excellent vision and immense knowledge. His guidance helped me in all the time of research and producing the original contributions that comprise this thesis. I could not have imagined having a better advisor and mentor for my Ph.D. study.

Besides my advisor, I would like to thank the rest of my examining committee members, including the external examiner, Dr. Dongmei Zhao of McMaster University, and the internal examiners, Dr. Yang Zhang and Dr. Pradeepa Yahampath, for the dedication of their invaluable time and expertise, and assessment of my thesis. Their insightful comments and suggestions have considerably improved the quality of this thesis. I am thankful to Dr. Douglas Thomson for serving as the chair of final Ph.D. oral examination.

I am grateful to my colleagues and fellow researchers in the Communication and Network Engineering Research (CNER) Group. Particularly, I would like to thank Dr. Xiaolong Li for his constructive discussions and excellent cooperation during his time at the University of Manitoba.

Last but not the least, I must express my very profound gratitude to my wife for her invaluable support, encouragement, and genuine dedications, along with my parents who are always behind me and have dedicated their lives to me. Without them, this accomplishment would not have been possible. I am truly lucky and thrilled to have them in my life.

This work was supported by the Natural Sciences and Engineering Research Council (NSERC) of Canada Discovery Grants.

This thesis is dedicated to my family.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background and Motivation

During the last decades, the demand for wireless communication services has dramatically increased, which includes not just telephony service but also video streaming and data applications. This trend in turn has triggered a wide deployment of wireless access networks. However, the power consumption in the Information and Communication Technology (ICT) industry also experiences a rapid growth, which has drawn a great attention to the public. To avoid the excessive greenhouse gas emission in ICT industry, it is imperative to control the power consumption, and meanwhile, to fulfill the ever growing users requirement and reduce the operational cost. There are also some problems in the traditional wireless networks that need to be considered to improve the network energy efficiency. First, the base stations (BSs) are commonly designed to satisfy the peak traffic requirements and always remain active status. Thus, it is viable to save energy by turning off some idle or under-

utilized BSs. Second, the BSs are operating statically and hence lack adaptability and cooperation. Third, the network traffic always changes with space and time and creating an unequally distributed traffic, which causes a problem of load imbalance. Finally, the introduction of small cells, such as femtocells (FCs), can enhance coverage to hotspots. However, the energy efficient operating still needs to be carefully investigated. Motivated by these problems, in this thesis, our objective is to develop the energy efficient strategies based on BSs on/off operation for green wireless networks. We first give an overview of the infrastructure in wireless networks and show the importance of green networking. Then, we introduce some potential techniques involved in green networking with the challenges and requirements.

1.1.1 The Importance of Green Wireless Networks

A wireless network is defined as a type of computer network that uses wireless data connections for connecting radio BSs or access points (APs) to mobile users (MUs) [1]. The wireless networks can include cellular networks, Wi-Fi networks, and terrestrial microwave networks, etc. A typical wireless network mainly consists of a core network, BSs/APs, and MUs [2]. The core network serves as a backbone network with Internet connectivity and provides data services. The BSs are responsible for radio resource management and user mobility management, and provide access to the Internet. In the hotspot area, multiple BSs may overlap or collocate to some extent. MUs can arbitrarily move, and at a given instant a particular MU can either be within or outside the coverage of BSs. The connection from the MU to a BS can be established by a single hop or using multiple hops when the MU is out of the

coverage of the corresponding BS.

In wireless networks, from the power consumption point of view, the BSs and the attributed back-haul networks consume approximately 60 billion kWh per year, which is corresponding to carbon dioxide emissions (CO_2-e) of 40 million tons per year or approximately annual gas emission of 8 million cars [3]. Among the devices of wireless networks, the BSs are responsible for 60% of the total power consumption of the entire networks [4]. In total, there are beyond 3 million BSs worldwide that consume 4.5 GW of power and cause approximately 20 Mt of CO_2-e in 2011 [5]. From the user side, it has been estimated that there are around 3 billion MUs in the world with power consumption of 0.2-0.4 GW [6]. Such high power consumption of wireless networks has promoted increased environmental and financial concerns for both service operators and users.

As increasing demand for more energy efficient technologies in wireless networks to tackle critical issues such as boosting cost of power consumption and excessive greenhouse gas emissions, the concept of green networking has drawn great attention in recent years. In fact, during the last decades, people have witnessed that carbon footprint of telecommunication industry has been exponentially growing due to explosive rise of service requirements and subscribers' demands. The concern on reducing power consumption raises from both environmental and economic reasons. From the environmental point of view, the ICT industry is responsible for approximately 2% of current global electricity demands with 6% yearly growth in ICT-related CO_2-e till 2020 [7]. For economics, the power consumption for operating a typical BS, which needs to be connected to the electrical grids, may cost approximately $3,000/year

while the off-grid BSs generally running on diesel power generators in remote areas may cost ten times more [8]. As more than 120,000 new BSs are deployed annually [9], it is still no end in sight on the development of mobile communications with many new subscribers and the constant desire of upgrading user equipment from 2G to 3G, then to 4G. The continuing growth in power consumption and carbon footprint of operating wireless networks has led to an emerging trend of addressing power-efficiency between the service providers and standard regulators. The term 'green' is widely mentioned in current research and the concept of green networking can be defined with various strategies and goals in different research fields. In wireless networks, the concepts of green networking [10] and green radio [11] were described from environmental, economic and regulatory points of view. In this thesis, for general telecommunication networks, we give our definition of green networking as "the implementation of energy-aware network technologies, protocols, and products, optimizing resources usage to reduce energy consumption with improved quality of service (QoS), and establishing next-generation economic and ecological telecommunication networks." The green networking can be achieved by implementing two strategies: *i*) matching the capacity offered by the network with the instantaneous traffic demand, which motivates the efforts in base stations (BSs) power control and management; *ii*) deploying small cells such as microcells, picocells, and FCs overlaid with existing macrocells to enhance the network capacity with fast data exchange and less power consumption.

1.1.2 BS Sleep Mode Techniques

At present, all BSs are working on the 'always-active state', regardless of the associated traffic level for each BS. In other words, a BS remains consuming power as usual when there is no traffic load in its coverage. Moreover, traditional BS deployment is designed to satisfy peak traffic requirements. In fact, the worldwide average peak utilization rates of the cellular networks are merely at 65% in 2011. Since the BSs are often underutilized in normal operation, it leaves a large room for saving network power consumption. This motivates extensive research on switching BSs between operational (active) mode and non-operational (sleep) mode[1] so as to achieve power saving concerning the fluctuations of traffic along space and time [12] [13]. In the standard of the 3rd Generation Partnership Project (3GPP) Long Term Evolution (LTE) Release 11 [14], it has been defined that the BSs can be switched off during the low traffic periods to save energy.

Many efforts on power saving have been done towards switching off some underutilized BSs. For instance, the authors in [13] investigated the feasibility of reducing the number of active BSs by considering the day-night behavior of mobile users. Some typical cellular network configuration such as crossroads, Manhattan, and hexagonal cell layout were considered. The authors concluded that switching off some underutilized BSs was possible to achieve large power saving. In [15], the authors developed dynamic BS power management for cellular networks and derived power saving ratio. The concept of cell zooming, which adjusted the cell size dynamically based on traffic variations was presented in [16]. The cell size could be reduced when the BS detects

[1]We use terms operational (active, power on) mode and non-operational (sleep/dormant, power off) mode to indicate BS working modes interchangeably in this chapter.

less traffic in its coverage while the adjacent cells enlarged their cell size to cover the gap correspondingly. In practice, cell zooming could be implemented by adjusting the physical parameters such as BS transmit power, or by the cooperation and relay among BSs. The authors in [17] demonstrated the insufficiency of cell zooming in some cases and discussed the feasibility of deploying smaller but more cells to increase energy efficiency. Another similar proposal on dynamically adjusting cell size in a multilayer cellular architecture, called 'cell breath', was introduced in [18]. Also, a decentralized BS sleeping algorithm was proposed for the LTE system, where the BS could be turned off when it had the smallest utility value based on the provided data rate for service and the maximum operational power [19]. In [20], the authors discussed the duty cycle of BS on/off and proposed an antenna selection criterion. A distance aware BS on/off scheme was proposed in [21], where the BS on/off control depended on the distances between BSs and the associated mobile users. In [22], Han et al. discussed a scalable BS switching scheme to extend network coverage to include service areas of inactive BSs. However, since the traditional BSs in macrocells ordinarily have wide coverages and serve many users, it's very complicated in practical implementation to switch on/off the macro-BSs by considering the challenges such as coverage holes and user associations.

1.1.3 The Characteristic of Heterogeneous Networks

The BS on/off operation has often been involved in small cells cooperation in heterogeneous networks (HetNets) since the coverage or the service of an inactive BS need to be taken care by the nearby BSs with joint work. The HetNets is a promising

Table 1.1: Comparison between power consumption and CO_2-e in base stations [24].

Cell Type	Radius (m)	Power (Wh)
Macro-cell	1000	31.4
Micro-cell	500	7.85
Pico-cell	100	0.31
Femto-cell	10	3.14×10^{-3}

Cell Type	CO_2-e (g)	Ratio (%)
Macro-cell	20.7	100
Micro-cell	5.18	25
Pico-cell	0.2	1
Femto-cell	2.07×10^{-3}	0.01

technique for wireless networks of next generation in response to the explosive growth on the request of faster wireless data services, more coverage as well as less cost on infrastructure construction and energy consumption. The HetNets consist of multi-tier cellular networks including macrocells, microcells, picocells, FCs and relay base stations. Such multi-tier network cooperations encourage the deployment of small-size cells, called small cell networks (SCNs), overlaid with the existing macrocell networks [23]. Due to their small coverage areas, SCNs require much less transmission power than the traditional macrocells. Table 1.1 shows the relationships between power consumption and the resultant CO_2-e by changing cell sizes. Obviously, the CO_2-e can be reduced significantly when deploying smaller cells.

Current wireless cellular networks are typically deployed homogeneously by using a macro-centric planning prototype. Although macrocells can provide larger coverage and better handle user mobility, they are not efficient in providing high data rates and can only work in the open access mode in the network (compared to the closed

access method with superior data rates provided by FCs) [25]. In addition, from the ecological point of view, the resultant CO_2-e of power consumption by varying cell sizes can be significantly reduced if deploying more small cells in wireless networks [26]. Moreover, the capacity limit of macrocells becomes more critical in dense urban areas, while on the opposite side, the deployment of small cells such as microcells and picocells can offer mobility and enhanced coverage to hotspots, and FCs can provide better indoor coverage. The SCNs are commonly required to be deployed in a quite dense manner with self-organizing, low-cost and low-power characteristics due to their small coverage radius. Such deployment of SCNs based on different types of cells has a potential to provide better service and increase the energy efficiency of wireless cellular networks. It also enables flexible and low-cost implementation and provides a uniform broadband experience to users anywhere in the networks.

Small Cells Cooperation

Currently, the evolution of wireless technology has made an isolated system (with a single type of BS) reach capacity limit determined by information theory [27]. Thus, further enhancement can only be achieved by developing advanced wireless topologies, one of which is the cooperated SCNs with self-organizing features. The main benefits of this solution are summarized as follows.

- The deployment of SCNs can eliminate dead spots in the conventional macrocell layout and improve network capacity in hotspots. In reality, there are always some areas with dense population, which may even require different types of service such as public, private and business services. There may also exist areas

with a sparse population where users may locate in coverage holes. Moreover, the wireless network traffic in some areas may be quite dynamic and unpredictable. Under all these scenarios, the cooperated SCN is an ideal solution to extend the coverage of the hotspots and uncovered areas with low implementation cost, adapt fast to the wireless traffic fluctuation and tend to provide various services.

- The deployment of small cells is more cost-effective compared to the traditional macrocells layout. The distributive placement of small cells can be based on a rough knowledge of coverage and traffic density, while the placement of macro BSs generally needs careful network-wide planning and centralized operation. In addition, the SCNs can implement the self-organizing network techniques [28], so that a smart network configuration in a cognitive approach with self-learning, intelligent decisions, and dynamic resource management can be adopted. The cooperation among small cells can also maximize the overall system capacity, coverage, and spectrum efficiency. Moreover, the self-organizing network enables the large deployment of small cells with economic viability by greatly reducing the costly human intervene in operation, administration and maintenance [29] [30].

- The deployment of SCNs can make the next generation wireless network more energy efficient, and contribute to the global energy conservation and CO_2-e reduction perspective. A small cell can be powered on or off more adaptively without worrying about the large cooling system as equipped in macro BSs. In addition, the small cells are likely to be turned off more often if there is

no associated traffic due to their relatively small coverage. Moreover, from the network point of view, the SCNs can operate cooperatively and adaptively by matching the wireless traffic variation, so as to selectively switch off some underutilized BSs during a certain period with the cells that remain on to take care of the coverage holes [13]. The number of switched-off small cells can be controllable based on the monitored network traffic levels. Such BS power on/off strategy in SCNs can significantly reduce the energy consumption of entire networks [16].

The deployment of a self-organizing network is also reflected in the current 3GPP standard (e.g., 3GPP TS 32.521). As shown in Fig. 1.1, the SCN aims to enable the network to optimize, reconfigure and recover itself so as to reduce energy expenditure and improve network performance and flexibility. For self-healing, the cells can detect coverage holes, diagnose defects, and make compensation by changing the cell coverage. For self-configuration, the cells can enable power control, access management, IP address and connectivity configuration. For self-optimization, the cells can perform load balancing, inter-cell interference coordination, coverage, and capacity optimization. In [23], the authors indicated cell size reduction was the simplest and most effective way to increase system capacity. They also discussed the challenges on deploying SCNs including self-organization, interference management, mobility, and security. In [29], the authors discussed offloading and distributed channel access techniques for the deployment of SCNs with the coexistence of macrocells, in which small cells served as offloading spots in the radio access network to offload users and their associated traffic from congested macrocells. It showed that small cells could overcome

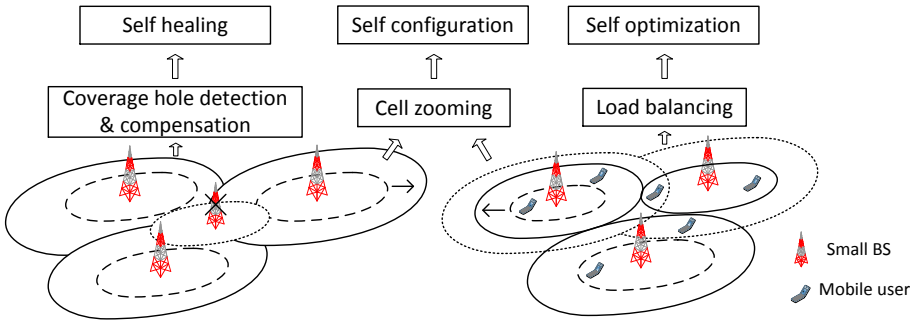


Figure 1.1: Main features of self organizing small cell networks.

the deployment challenges and efficiently coexist in a multi-tier cellular network with cognition capabilities (e.g., spectrum sensing). In [31], the authors indicated that self-organizing was effectively the only viable way to achieve optimal performance in future wireless networks in a cost effective manner. Also, the main objectives of self-organizing networks towards SCN perspective could be summarized as coverage expansion, capacity optimization, energy efficiency and QoS enhancement.

The evolution of self-organizing SCNs motivates us from two aspects. First, the current green networking strategy pays little attention to self-organizing SCNs. In fact, the small cells configuration can be more adaptable and manageable regarding network traffic distribution. Second, most existing works considering BS sleep mode adopted a unified BS-off strategy over the entire cellular network, and missed the consideration on the potential heterogeneity of traffic distribution among different zones. The zones with low traffic can actually use more aggressive BS-off strategy than the hotspots so that further improvement on power saving can be achieved.

Load Balancing and User Association

In HetNets, since there are many types of small BSs (SBSs) coexisting and overlapping with complex user traffic fluctuations, it is often very complicated to determine which small cells should be switched on and off to hold a balancing load. A typical challenging situation is that when user traffic dramatically increases in the coverage of a macro base station (MBS), this overloaded MBS may request wake-up of one or more dormant SBSs to offload some traffic, which however, may not be the most appropriate SBSs to turn on from energy efficient perspective. In addition, when the network has low traffic, it is also difficult to decide which SBS should be kept active if there are multiple choices to switch off some SBSs. In summary, it is essential to investigate an energy efficient SBS on/off strategy by jointly considering load balancing and interference coordination in HetNets.

In HetNets, another big challenge is the user association problem, which refers to dynamically assign mobile users to appropriate BSs so as to achieve load balancing and avoid making some cells overloaded. Traditionally in wireless networks, the most obvious method is max-SINR user association, which is simply assigning each user to the BS with highest received signal strength. However, simulations and field trials have shown that such an approach cannot increase the overall throughput as much as expected. Due to much stronger transmission power from the MBS against the smaller ones, most users are finally served by the MBS while many small cells are sitting idle [32]. This motivates a method called biasing user association, where users were actively assigned to nearby small cells [33]. Such strategy had the potential for a win-win solution because the mobile users gained access to a much larger fraction

of the small cell's time and frequency slots. Instead, the macrocell could reclaim spectrum that users would have occupied and reassigned it efficiently.

In multi-tier HetNets, the biasing user association motivates to re-evaluate the rules of user association and load balancing developed for tradition macrocell-only networks. In particular, it is currently unclear how much biasing is "optimal", which can be determined by many factors: (i) the throughput/QoS metric of interest, such as the energy efficiency and load balance, (ii) the trade-off among the performance metric, interference coordination and BS on/off operation, and (iii) the traffic pattern of mobile users and how various base stations are distributed in space.

1.1.4 Challenges in the Operation of Femtocells

Among various types of small cells, the FCs were introduced for future wireless networks driven by the demand of high data rate and enhanced energy efficiency [34]. In an FC, the BS or access point, also known as Home Evolved Node B (HeNB), is typically a consumer-deployed cellular access point installed by a subscriber in a home or SOHO (small office/home office)[2], which can provide better link quality for home users with low transmit power and can offload traffic from macrocells. In the standard of 3GPP Release 9 [35], a novel hybrid access concept was introduced for the HeNB, which required service provisioning to both open access users and closed subscriber group (CSG) users. For the hybrid access, mobile users belonging to the CSG received preferential service over unsubscribed users from open access.

In [36], the authors proposed optimal sleep and wake up schemes for macro-femto

[2]Femtocells can also be deployed by service providers based on its functions, such as capacity enhancement for hotspots and coverage extension to indoor and underground.

heterogeneous networks, where the FCs worked in open access mode and could offload traffic from the macrocells. The basic idea was that when the macro-femto heterogeneous network was lightly loaded so that the macro-BS could handle all the traffic alone, the FCs were switched off. Vereecken et al. in [37] proposed heuristic BS on/off decisions based on the number of user connections to different BSs. However, the strategy assumed that both users and BSs' locations were known. Estrada et al. in [38] studied optimal resource allocation in two-tier networks with hybrid access FCs. The authors claimed that effective spatial reuse between the macrocells and FCs could be achieved by joint power control in both tiers.

While unlike the previous research which mainly addressed the joint operations between the macrocells and FCs, other studies in [39–41] focused on the operations of the FCs only. In [39], Li et al. proposed a simple sleeping scheme, called fixed time sleeping, for saving power of the FCs. However, the fixed time sleeping scheme may not be well adapted to the traffic variations in practical networks. Ge et al. in [40] conducted performance analysis for two-tier FC networks with open access users. Based on derived Markov chain models, the spectrum and energy efficiency were analyzed under different scenarios in terms of the number of FCs, the average number of users, and the number of open channels. Kim et al. in [41] studied the effects of the FCs sleeping ratio on the energy efficiency by using a stochastic geometry-based model and derived the optimal sleeping ratio to maximize the energy efficiency. However, neither [40] nor [41] addressed the issues of the effects of users priorities or the constraints on users delay and blocking rates.

Although existing studies considering the HeNB power on/off have successfully

enhanced the system energy saving, these studies have not jointly considered the hybrid access and the HeNB power on/off characteristics, which are essential for practical implementations. By considering user priorities, e.g., when the high priority user from CSG can interrupt the HeNB's sleep periods, the HeNB on/off scheduling becomes more complicated in order to balance the energy saving and QoS provisioning in terms of the queue lengths, the user waiting time and blocking rates for different types of users.

1.2 Contributions

The objective of this research is to develop the energy efficient strategies based on BS on/off power management and the characteristics of small cells, so as to bring reliable and effective green wireless networks. The main contributions of this thesis are summarized as follows.

- We first address the issue that traditional unified BS-off method is not able to transform effectively with the traffic fluctuations in time and space. We proposed two energy efficient schemes: the centralized clustering BS off (C-CBSO) protocol with the aid of BS controller, and the distributed clustering BS off (D-CBSO) protocol with the collaboration of BSs. Both of them adaptively adjust and form BS clusters, which later run the optimal BS-off scheme individually based on the traffic distribution in order to switch off the underutilized BSs as many as possible. Experimental results testify that the proposed methods can save more than 50% energy cost of the entire cellular network and 15.1% more energy efficient compared to the unified BS off method. The detailed contents

are shown in Chapter 2.

- We further develop a distributed small cell management strategy, called self-organized BS virtual small network (VSN) protocol to reduce power consumption in wireless networks. The virtual small networks are formed based on the heterogeneous traffic distribution. In each VSN, a number of the underutilized BSs can be switched off during a certain period. To deal with the VSN transformation problem, we introduce the *reception*, *analysis* and *control* (RAC) process. In *reception* stage, each BS applies the fuzzy inference system (FIS) algorithm to calculate the realistic traffic level based on the collected information including the number of MUs, the average distance between MUs and BS, and the timeline. In the following *analysis* stage, by utilizing hidden Markov model (HMM), the BSs estimate the mean network traffic distributively by applying collected traffic information from the neighboring BSs. In the final *control* stage, we propose the BSs of VSN with firefly (BSVF) algorithm, which targets to adaptively group BSs based on information observed from previous stages in each operation period. Instead of running a unified BS-off strategy in the entire network, the modified BS-off matching scheme is applied to make a selection of the best BS-off strategy for each VSN separately. Moreover, the BS on/off weight is introduced as a learning process to revise the probability of BS modes transition, which can avoid the BSs making frequent transitions between active and sleep modes. The detailed contents are shown in Chapter 3.

- We also consider to improve the energy efficiency of the HeNB. To capture the transmitter power on/of and hybrid access behavior of the HeNB, we formulate

an MAP/PH/1/k queue to model the HeNB with multiple vacations, exhaustive service discipline, and non-preemptive priority. Then we derive the key performance metrics resulting from queue length, user waiting time, blocking probability and system busy cycle. In contrast to the existing literature [42–44], one of the major differences in our MAP/PH/1/k queueing analysis is that the user priority and multiple vacations are jointly considered in order to better match the HeNB characteristics In addition, the proposed queueing model also has the vacation termination policy with a preference to high priority users. The detailed contents are shown in Chapter 4 Section 4.2.

- To address the problem of maximizing HeNB energy efficiency by considering power on/off mechanism and user priorities and satisfying the QoS requirements, we further propose an adaptive service rate and vacation length (ASV) method. The optimization problem is then solved by the Lagrange decomposition method, where the adaptive service rate is derived in the inner optimization, and the adaptive vacation length is derived in the outer optimization. A one-step look-ahead method is proposed in order to reduce the computational complexity of the ASV method. The one-step look-ahead method is based only on buffer size and the number of admissible service rates in the current stage, so that it is more computational effective compared to the ASV method. Extensive simulation results are provided to demonstrate the efficiency and effectiveness of the proposed adaptive service rate and vacation length strategy. Given a buffer size, the proposed ASV method outperforms the traditional one with fixed service rate and vacation length by up to 20% in terms of system energy

efficiency. Such gain can be further enhanced when the HeNB adopts a larger buffer size. The detailed contents are shown in Chapter 4 Section 4.3.3.

The rest of this thesis is organized as follows. The detail of the proposed centralized and distributed clustering BS off schemes for the green wireless network is described in Chapter 2. In Chapter 3, the novel self-organized virtual small networking protocol for power saving and load balancing of small cell networks are investigated. The strategy of adaptive service rate and vacation length for energy-efficient HeNB based on MAP/PH/1/K queueing analysis is studied in Chapter 4, followed by conclusions and future work in Chapter 5.

Chapter 2

Energy-efficient Base Station Control with Dynamic Clustering

In this chapter, the BSs on/off and BSs clustering strategy is discussed. First, the system model is defined. Then, both centralized and distributed CBSO protocols are proposed, as well as the formulated BS-off matching scheme so as to increase the power saving of the entire network. Finally, the simulation results are presented in the last section.

2.1 System Model

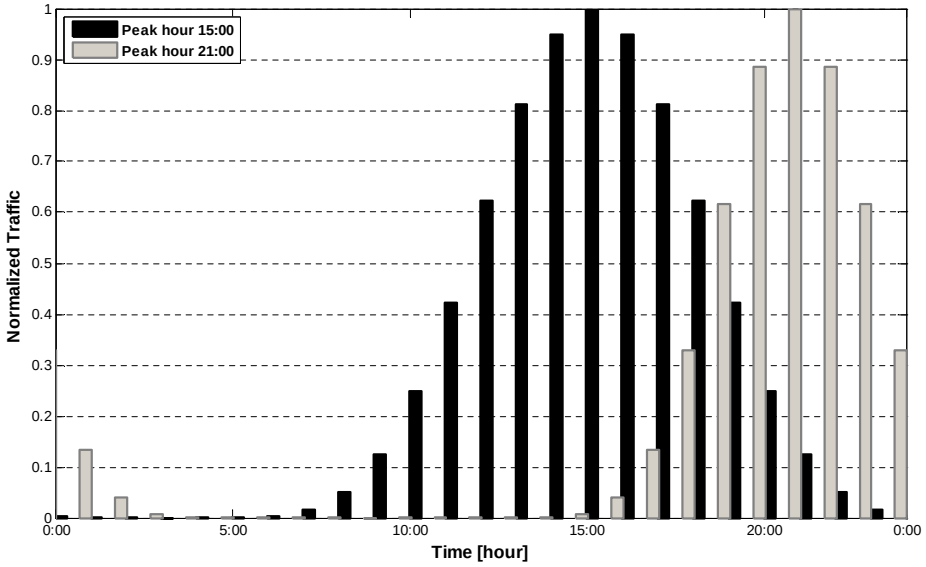
Consider a cellular network deployed in an urban scenario, which is covered by homogeneous micro hexagonal cells. Each cell is covered by a BS located in the geometric center. All the cells may experience various traffic levels over space and time as shown in Fig. 2.1. For example, peak traffic hours always appear in office

areas in the daytime, while high traffic periods usually come from residential areas at night. All the BSs use omni-directional antennas. We say BS i is BS j's adjacent BS, if BS i's covering region is adjacent or overlapping with BS j's one. We assume that when a BS is switched off, its adjacent BSs that remain on can provide service to the coverage gap area. By following the similar discussions on microcells as shown in [13] [45], the power increment due to the coverage enlargement is omitted in this chapter[1].

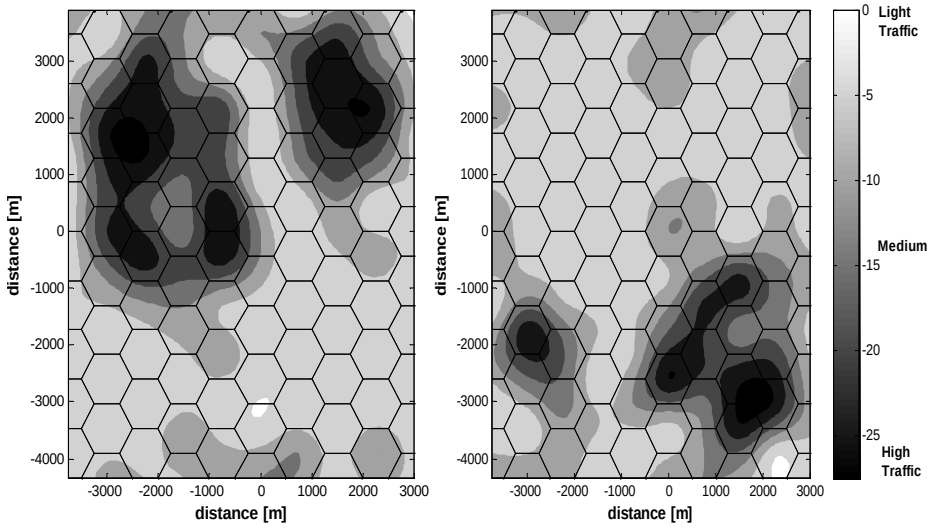
2.1.1 BS Power Consumptions

The traditional power consumption of a BS is formulated as $P_{BS} = P_{sta} + P_{dyn}$ [46], where P_{sta} is the static power consumption and denotes the power consumed when a BS is idle or does not communicate with any mobile users (MUs). P_{sta} includes power consumption by the power amplifier (PA), feeder, power supply, signal processing, and air conditioning. P_{dyn} indicates the dynamic power consumption, which mainly determined by the power used for data transmission. Obviously, P_{dyn} depends on the associated traffic load. In this chapter, we also take into consideration the transition cost of a BS, P_{tc}, which means the extra power consumption in transmitter transit period (when the working status of BS transforms between the active and sleep modes) defined in the standard [14]. P_{tc} is calculated as the percentage of BS's static power consumption and is formulated as $P_{tc} = \rho P_{sta}$, where ρ is the transition cost ratio. In

[1]In urban UMTS scenarios, in the few cases, the emitted power should be increased when extending the cell coverage. Such power increase can be negligible (a few watts) compared to the total BS power consumption (more than a hundred watts).



(a)



(b)

Figure 2.1: (a) The traffic distributions changing with time in distinct hotspots. (b)The traffic distributions in space at observation time 15:00 and 21:00 respectively.

summary, we reformulate the power consumption of a BS as

$$\begin{aligned} P_{BS} &= P_{sta} + P_{dyn} + P_{tc} \\ &= (1+\rho)P_{sta} + P_{dyn}. \end{aligned} \tag{2.1}$$

It is assumed the power consumption on computation is purely in software and can be negligible compared to the total BS power consumption.

2.1.2 Traffic Model

We revise the traffic model in [47] to better match the reality. The revision is based on the fact that most BSs do not experience maximum traffic load in a day, especially those BSs locating outside the hot zones. Moreover, the microcell deployment implies that the number of MUs served by each BS is small, which leads to more random traffic variations among cells. Thus, we reformat the traffic model in a cell as,

$$f(t) = \frac{\mu}{2^{\nu}}[1 + \sin(\pi t/12 + \varphi)]^{\nu} + \sigma(t), \tag{2.2}$$

where $f(t)$ is the normalized traffic fluctuating with time t. μ is a uniformly distributed random variable in the interval [0, 1], which controls the peak traffic rate in a microcell. $\nu = 1, 3, \text{or } 5$, which determines the abruptness of the traffic model. A larger ν means that the traffic curve has a steeper slope while the average traffic load is lower. φ is a uniformly distributed random variable in the interval $[\frac{3\pi}{4}, \frac{7\pi}{4}]$, which determines the distribution of traffic with different peak hours among BSs. $\sigma(t)$ is a Poisson distributed random process, which is used to model the random fluctuations of the total traffic. Note that the modified traffic model can restrict the peak hours in a rational period instead of a whole day.

2.2 BS Clustering Protocols

Following the motivation to operate distinct BS-off schemes in different traffic zones, the entire network is partitioned into several clusters and such cluster arrangement may change with time and space. At each observation instant, the network runs the proposed clustering BS-off (CBSO) protocols.

In centralized CBSO (C-CBSO) scheme, the base station controller (BSC) gathers all the information, analyzes and manages the cluster formation and selects the cluster head. While in distributed CBSO (D-CBSO) algorithm, each BS shares the information only with its adjacent BSs, makes its own decision to join a cluster and selects cluster head. Both proposed CBSO protocols with either centralized or distributed are able to operate dynamically and adaptively without human intervene. Moreover, the BS-off matching scheme is developed to pick the most feasible BS-off scheme for each cluster.

2.2.1 Centralized CBSO Protocol

Consider a cellular network with J BSs forming a grand coalition N. In centralized cluster formation, the BSC first initializes total K clusters as S_k, $k = \{1, 2, ... K\}$, where $K > 1$ is mandatory. To better apply BS-off matching schemes and simplify the cluster forming process, K should be set as a small number. For example, K can be determined based on the number of hot zones in the network. If there are total Z hot zones, the initial number of clusters can be set as $K = Z + 1$. Each cluster has n_k BSs and the cluster head is denoted by H_k. Obviously, we have $N = \underset{k}{\cup} S_k$. At the initial stage, the cluster head, H_k, is selected by the BSC so that H_k can be located

in the center of S_k. The BSC also maintains a fringe BS set F. BS j belonging to F means that BS j and at least one of j's adjacent BSs are associated with different clusters. The formulation of F is to guarantee that only BSs in F are allowed to join one of the adjacent clusters during each operation period[2], and a non-fringe BS is never separated from its associated cluster. In addition, define C_j as a set of BS j's adjacent clusters, which means that BS j has opportunities to join one of the clusters in C_j if $j \in F$.

In C-CBSO process, the BSC first broadcasts a message to all BSs and calls for data including BSs location information, adjacent BSs list, and timely traffic statistics. After that, the BSC calculates BS j's, $j = 1, 2, \ldots, J$, power consumption $P_j(t)$ and compare $P_j(t)$ with the mean power of C_j at the observation time t by using following utility function

$$U_{j,k}(t) \quad = \quad |P_j(t) - \bar{P}_{j,k}(t)| \ , \quad j \in F,\ k \in C_j, \tag{2.3}$$

where $\bar{P}_{j,k}$ represents the average power of C_j and can be formulated by

$$\bar{P}_{j,k}(t) = \frac{\sum_{i=1}^{n_k} P_i(t) - \omega_{j,k}(t) P_j(t)}{n_k - \omega_{j,k}(t)}, \quad j \in F,\ k \in C_j, \tag{2.4}$$

where $\omega_{j,k}(t) \in \{0, 1\}$ is a binary variable, which stands for the associated relationship between BS j and cluster S_k. If $\omega_{j,k}(t) = 1$, it means BS j is a member of cluster S_k at instant t. Otherwise, $\omega_{j,k}(t) = 0$. Since each BS has only one associated cluster, we have

$$\sum_{k=1}^{K} \omega_{j,k}(t) = 1, \quad \forall j = 1,\ 2,\ \ldots J. \tag{2.5}$$

[2]The operation period is a predefined parameter by service's provider in hourly scale, e.g., 2, 3 or 4 hours.

After that, the BSC notifies BS j to join a cluster S_k with the minimum $U_{j,k}(t)$. During each operation period, any BS can only change its associated cluster no more than once. After cluster formation, BSC provides processed data to cluster heads in order to help them run the proposed BS-off matching scheme as introduced in Section 2.2.3. In addition, F should be reset when the cluster formation process is completed. Since the static power consumption P_{sta} is identical for all BSs, $P_j(t)$ is mainly determined by the number of the associated users in each BS. Thus, the cluster formation process actually groups the BSs with similar traffic load together.

Algorithm 1 Distributed CBSO operation

Input: J, K, T, F, S_k, H_k, $Q_{j,k}$.
Output: F, S_k, H_k, $Q_{j,k}$.
Begin
foreach $t \in T$
  while $F \neq \emptyset$ **do**
    foreach $j \in F$.
      compute power consumption $P_{j,k}(t)$
      update cluster association variable $w'_{j,k}(t)$ of BS j
      compute the estimated average power consumption
        of j's adjacent cluster by using 2.6
      find $\min |P_{j,k}(t) - \bar{P}_{j,k}(t)|$
      set $S_k \leftarrow S_k + \{j\}$, $F \leftarrow F - \{j\}$
  end while
  foreach $k \in K$
    foreach $j \in S_k$
      set $Q_{j,k} \leftarrow Q_{j,k} - 1$
      if $Q_{j,k} = 0$, **then**
        select BS j as new cluster head H_k of S_k
      end if
    run BS-off matching scheme at H_k
  update F
End

2.2.2 Distributed CBSO Protocol

For self-organized cluster formation, since no global information is available, each BS determines its cluster association merely depending on the statistic data gathered from its adjacent BSs. At the initial step, the cluster head, H_k, is randomly selected and each BS joins S_k if it has the shortest Euclidean distance to H_k. Each BS shall maintain a binary matrix, $w'_{j,k}$, $j = \{1, 2, ... J_{nb}\}$, which indicates the cluster association status of its adjacent BSs. $w'_{j,k} = 1$ means the adjacent BS j is the member of cluster S_k, and $J_{nb} = 6$ for hexagonal cell layout. Also, a BS shall judge based on $w'_{j,k}$ that whether it is currently a fringe BS or not (the BS located at the edge of a cluster). If a BS j is associated with cluster S_k, and $\sum_{j=1}^{J_{nb}} w'_{j,k} < J_{nb}$, BS j is said to be a fringe BS, i.e., $j \in F$.

In D-CBSO precess, BS $j \in F$ collects and shares data with its adjacent BSs, including traffic loads, power consumption and clustering associated status. The estimated mean power consumption $\bar{P}_{j,k}(t)$ of a cluster S_k at BS j is calculated by

$$\bar{P}_{j,k}(t) = \frac{\sum_{j=1}^{J_{nb}} w'_{j,k}(t) \cdot P_j(t)}{\sum_{j=1}^{J_{nb}} w'_{j,k}(t)}, \quad j \in F. \tag{2.6}$$

Note that all the information from the adjacent BSs is considered. Then, BS j joins cluster S_k when it has minimum $U_{j,k}(t) = |P_j(t) - \bar{P}_{j,k}(t)|$. After that, the fringe BS set F is updated, and the cluster heads H_i are selected again due to the update of BSs in each cluster. H_k will then run the BS-off matching scheme in cluster S_k. To ease the cluster head selection, each BS maintains a back counter number $Q_{j,k}$ with distinct values, which are positive integer randomly initialized when a cluster is formed and decreased after each cluster formation process. The BS is selected to be the cluster head when its $Q_{j,k}$ reduces to 0. At the end of each operation period,

the new cluster head broadcasts and sets up connections with its cluster members. Algorithm 1 summarizes the details of the D-CBSO.

Compared to the C-CBSO algorithm, the proposed D-CBSO algorithm leads to no information congestion at BSC and lightens the information burden at cluster heads. The mitigation of information exchange among BSs and cluster heads may result in additional power saving. Moreover, a much faster process of cluster formation becomes possible.

2.2.3 BS-off Matching Scheme

After cluster formation process, each cluster head needs to choose the best BS-off scheme for its own cluster. The selection process can be formulated to an optimization problem as follows. For any cluster, let B denote a set of all available BS-off schemes. For explanation purpose, we adopt the widely used (n, m)-off strategy [13] [48], which switches n out of m active BSs to sleep mode. The optimization problem to minimize power consumption of the network, P_{nw}, is formulated as

$$\begin{aligned}
\min P_{nw} =& \sum_{t} \sum_{k \in K} \sum_{j \in n_{k,a}} P_{j,k}^{(B_k)}(t) \\
=& \sum_{t} \sum_{k \in K} \sum_{j \in n_{k,a}} \left((1+\rho) P_{sta}^{(B_k)} + \left[P_{dyn}^{(B_k)}(t) \right]_{j,k} \right) \qquad (2.7) \\
=& \sum_{t} \sum_{k \in K} \sum_{j \in n_{k,a}} \left((1+\rho) P_{sta}^{(B_k)} + \eta \cdot w_{j,k}(t) \cdot l_{j,k}^{(B_k)}(t) \right) \\
\text{s.t.} \quad & \sum_{i \in m} l_{i,k}^{(B_k)}(t) \leq (m-n) l_{max},\ i = 2, 3, ..., J_{nb}+1 \qquad (2.8) \\
& m \leq n_{k,a} \leq n_k,\ m = 2, 3, ..., J_{nb}+1 \qquad (2.9)
\end{aligned}$$

where $P_{j,k}^{(B_k)}(t)$ is the power consumption of BS j in cluster S_k at observation time t under B_k-th feasible (n, m)-off scheme. $n_{k,a}$ indicates the number of active BSs in cluster S_k. η stands for a coefficient of BS dynamic power consumption P_{dyn}. $l_{j,k}(t)$ describes the instantaneous traffic load of BS j. Constraint (2.8) guarantees the remaining active BSs can maintain the service of sleep BSs, while the aggregated traffic loads of both active BSs and the covered sleep BSs cannot exceed overall predefined maximum user capacity l_{max}. Since (n, m)-off scheme is applied in each cluster, constraint (2.9) means that the number of active BSs, $n_{k,a}$, in cluster S_k has to be larger than m. In practice, since the set of feasible BS-off strategies B is finite, exclusive search can be adopted in all potential candidates to find the best one for implementation[3].

2.3 Simulation Results

2.3.1 Simulation Setup

In our simulation, the cellular network consists of total $J = 100$ BSs with the deployment of a micro hexagonal cell layout. The network covers approximately 54 km^2, which is equivalent to a small town region. The minimum cell radius is set to 500 m with BSs transmission range of 866 m, which is corresponding to $P_{sta} = 237$ watt for a typical BS. BSs have 10 watt power consumption when they are turned to sleep mode. For simplicity, the total number of hotspots is set to 4 with abruptness factor $\nu = 3, 5$ in (2.2), while other fields with $\nu = 1$ represent the low traffic zones.

[3]The feasible BS-off strategies are a finite set, since in hexagonal cell layout, in (v, u)-off scheme, we have $v < u \leqslant 7$.

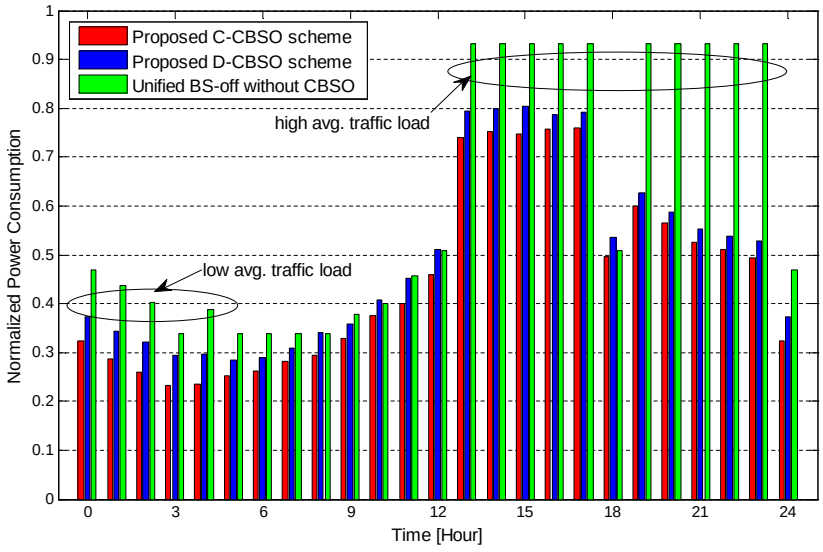


Figure 2.2: Power consumption of the proposed CBSO schemes.

The arrival process of MUs follows Poisson distribution. The traffic loads fluctuate with time and randomly generate in both hot zones and low traffic areas, which approximates a real urban traffic model. Fig. 2.1 shows an example of the traffic distributions in two specific hours during a day.

The path loss model is compliant with microcell test environment in ITU report with a center frequency of 2.655 GHz [49] and the receiver sensitivity of MUs is -120 dB. The power supply loss is approximate 10% and the efficiency of PA is 20%. The maximum number of MUs can be served by a BS is $l_{max} = 25$, which corresponds to assigning 5 MHz bandwidth in an LTE system.

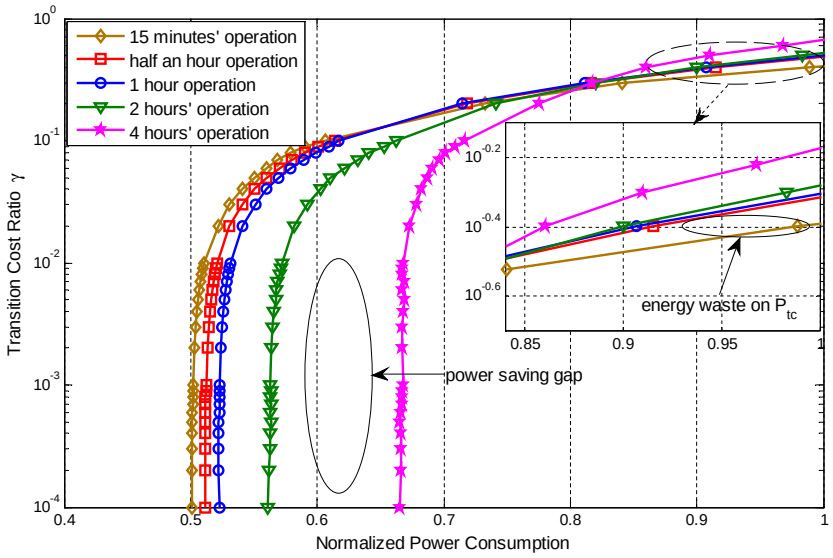


Figure 2.3: Power consumption rates with different operation intervals.

2.3.2 Performance Evaluation

Fig. 2.2 shows the comparison of power consumption between the proposed two CBSO schemes and traditional unified BS-off strategy [13] [48]. The simulation results are normalized with respect to the network power consumption without any cluster formation and BS-off mechanisms. It shows that an average 50.4% power consumption can be saved for D-CBSO scheme compared with that of 54.9% for C-CBSO scheme. It is obvious that C-CBSO outperforms D-CBSO since more information is available for BSC to carry out better cluster formation and BS-off scheme selection. Moreover, it is manifest that our proposed schemes are able to gain 16.1%-19.6% more power savings than the unified BS-off strategy. The reason is that the proposed CBSO schemes are capable of balancing the traffic load in each cluster when the network traffic load of certain time and area is quite high or low.

Fig. 2.3 shows normalized power consumption under diverse operation intervals with the change of transition cost ratio ρ. It shows that when ρ is small, i.e., the transition cost can be negligible, a shorter operation period is better since the protocol can better match the network traffic fluctuation. However, when ρ is large, too often operation leads to more power waste. As shown in Fig. 2.3, the reasonable and applicable operation intervals turn out to be from half an hour to two hours.

Chapter 3

Self-organized Virtual Small Networking Protocol for Energy Saving and Load Balancing in Wireless Networks

In this chapter, the concept of BS virtual small networks (VSNs) is proposed, which can be self-organized adaptively based on the cooperation of small cells, powering off the underutilized small cells by monitoring the current traffic distribution, and guaranteeing the required QoS at the same time. The remainder of this chapter is organized as follows. We first describe the general system model of wireless VSNs. Then, the methodologies of determining BS realistic traffic level and estimating mean network traffic are investigated. After that, we propose the self-organized BS VSN firefly (BSVF) algorithm, as well as the BS-off matching scheme and BS

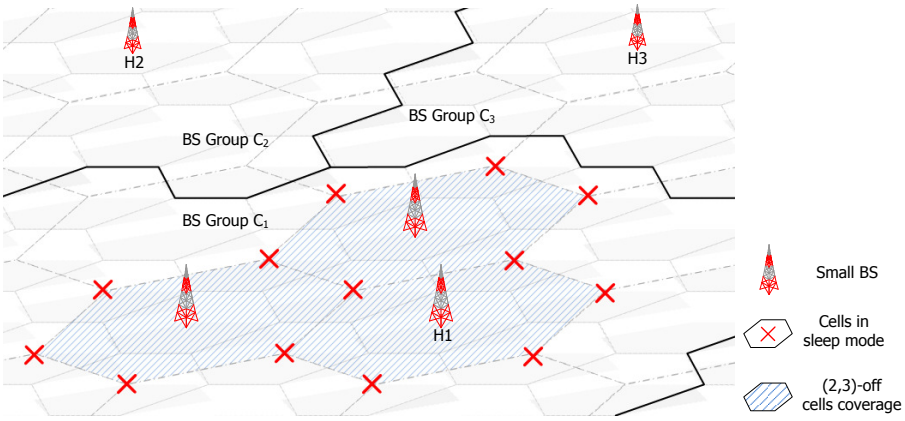


Figure 3.1: The system model of wireless networks with the deployment of homogeneous microcells. (It consists of BS groups $\mathcal{C}_1, \mathcal{C}_2, \mathcal{C}_3$ with BS headers $\mathcal{H}_1, \mathcal{H}_2, \mathcal{H}_3$, respectively. A $(2, 3)$-off strategy is adopted in $\mathcal{C}_1$, in which the power consumption can be reduced by switching off 2 BSs in every 3 BSs.)

on/off weight. Finally, the simulation results in terms of power consumption, CO_2-e, and convergence rate on VSNs' average traffic are presented.

3.1 System Model

We consider an infrastructure-based cellular network in an urban environment consisting of multiple micro BSs. Let $\mathcal{N}$ denote a grand coalition of homogeneous micro hexagonal cells serving this urban area. Each cell with radius $\mathcal{R}$ is covered by one micro BS located at the center and all BSs are equipped with omnidirectional antennas. The Euclidean distances, r_{ij}, of any two neighboring BSs i and j are identical.

Definition 1 (neighboring BSs)**.** BS i is the neighboring BS of BS j if BS i and BS j have adjacent geographical covering regions. In hexagonal cells layout of cellular networks, we have $r_{ij} = \sqrt{3}\mathcal{R}$.

Define n_j as the number of neighboring BSs of BS j. Since we consider identical micro hexagonal cells layout, we have $n_j^{max} = 6$. In the cellular network, each BS and its neighboring BS have opportunities to be grouped into a cell set $\mathcal{C}$, each of which is called virtual small network (VSN). At the initial stage, the total number of VSNs is assumed to be K and any VSN in the network is randomly formed. The detail on determining K is provided in Section 3.3.4. Each VSN, $\mathcal{C}_k$, consists of c_k, $k = \{1, 2, ..., K\}$, neighboring BSs. Thus, we have $\mathcal{N} = \cup_{k \in \mathcal{K}} \mathcal{C}_k$. When the VSNs are set up, the BS located at the center of each VSN is initialized to be the VSN controller, $\mathcal{H}_k$, whose task is to collect data from member BSs within the same VSN and coordinate collaboration among all member BSs. To reflect the practical urban environment, each BS i may experience different traffic densities changing with space and time. For example, peak traffic hours usually appear in office districts in a daytime or residential regions at night. For any VSN, define Ω as a set of all available BS-off strategies. For explanation purpose, we adopt the widely used (v, u)-off strategies [13] [48] as an example. An (v, u)-off strategy means switching v out of u active BSs to sleep mode, where $v < u$ and both u, v are positive integers. The system model is shown in Fig. 3.1 and illustrated with an example of $(2, 3)$-off strategy. In this chapter, we further introduce following three requirements on implementation:

Remark 1: The radio coverage and service of BSs in sleep mode can be maintained by the remaining active BSs, i.e., the QoS of MUs has to be guaranteed in dormant cells. The power increment due to the coverage enlargement is omitted, which follows the similar discussions on microcell layout as shown in [45].

Remark 2: The entire cellular network is considered in a distributed manner, in

which direct information exchange is not allowed between BS i and BS j if they belong to different VSNs. It means the entire network condition is unknown for an individual BS. However, the neighboring BSs can exchange information such as traffic level, BS power on/off states and VSN association status.

Remark 3: The traffic distribution has the similar profiles in daily measurement. Hence, the network power consumption is measured in a daily interval and the time is divided into fixed-length intervals as time slots.

The power consumption of a micro BS follows the similar discussion in Chapter 2, which consists of static and dynamic power consumptions, denoted by $\mathcal{P}_{sta}$ and $\mathcal{P}_{dyn}$, respectively. $\mathcal{P}_{sta}$ mainly consists of power cost by the power amplifier (PA), power supply, air conditioning, etc. $\mathcal{P}_{dyn}$ indicates the dynamic power consumption, which represents the power cost for data transmission. Obviously, $\mathcal{P}_{dyn}$ depends only on the traffic load of a BS. We also take into consideration the power cost in the transmitter transient period, when the BS working states transform between the active and sleep modes, denoted as $\mathcal{P}_{tc}$, as defined in LTE 3GPP Release 11 [14]. It is essential to prevent frequent transition between different BS working modes. In summary, the power consumption model of BS i can be formulated as

$$\mathcal{P}_i = (1 + \rho)\,\mathcal{P}_{sta} + \mathcal{P}_{dyn}, \tag{3.1}$$

where $\mathcal{P}_{tc}$ is a constant and is derived as a ratio of BS static power consumption, i.e., $\mathcal{P}_{tc} = \rho\mathcal{P}_{sta}$, where ρ denotes the power ratio of transmitter transient cost, and $\rho = 0$ if the BS working mode is unchanged.

3.2 Problem Formulation

Our objective is to minimize total power consumption of the entire cellular network by designing the VSN formation protocol with diversified (v, u)-off strategy. The optimization procedure is considered within a certain period of time $\mathcal{T}$, e.g., a hourly based measurement. Also, let T denote the total number of time slots. The optimization problem can be formulated as

$$\begin{aligned}
\min_{\mathrm{c_k},\Omega_\mathrm{k}} \mathcal{P}_{nw} &= \int_0^{\mathcal{T}} \left(\sum_{j=1}^{N} \mathcal{P}_j \right) dt \\
&= \sum_{t=0}^{T} \sum_{k=1}^{K} \sum_{j=1}^{c_k} \mathcal{P}_{j,k}^{(\Omega_k)}(t) \cdot \omega_{j,k}(t) && (3.2) \\
\text{s.t.} \quad & \sum_{k=1}^{K} c_k = N && (3.3) \\
& \mathcal{P}_{j,k}^{(\Omega_k)}(t) \geq 0, \forall j \in \mathcal{N}, \forall k \in \mathcal{K} && (3.4) \\
& \sum_{k=1}^{K} \omega_{j,k}(t) = 1, \forall j \in \mathcal{N} && (3.5) \\
& \sum_{j=1}^{N} \omega_{j,k}(t) = c_k, \forall k \in \mathcal{K} && (3.6) \\
& \sum_{i=1}^{n_j} \omega_{i,k}(t) \leqslant n_j^{max}, \text{when}\, \omega_{j,k} = 1 && (3.7) \\
& v < u \leq c_k, \forall k \in \mathcal{K}, (v, u) \in \Omega_k && (3.8) \\
& \sum_{j=1}^{u^{(\Omega_k)}} l_{j,k}(t) \leq (u - v)^{(\Omega_k)} l_{max}, \forall k \in \mathcal{K} && (3.9)
\end{aligned}$$

where $\mathcal{P}_j$ denotes the power consumed by BS j. $\mathcal{P}_{j,k}^{(\Omega_k)}$ represents the power consumption of BS j in VSN $\mathcal{C}_k$ under the BS-off strategy Ω_k. The binary value $\omega_{j,k} \in (0, 1)$ denotes the association status between BS j and VSN $\mathcal{C}_k$. If $\omega_{j,k} = 1$, it means BS j is associated to VSN $\mathcal{C}_k$. Otherwise, $\omega_{j,k} = 0$. Constraint (3.3) satisfies the total

number of BSs in all VSNs and that in grand coalition $\mathcal{N}$ are equal. Constraint (3.4) limits the power consumption of BS to be non-negative values. Constraint (3.5) means that a BS is associated with only one VSN at any time t. Constraint (3.6) shows the calculation of c_k. If BS j is a member of VSN $\mathcal{C}_k$, constraint (3.7) limits the number of BS j's neighboring BSs, n_j, which are associated with VSN $\mathcal{C}_k$, cannot exceed n_j^{max}. Since different BS-off strategy is applied in each VSN, and there are v in u BSs to be powered off, constraint (3.8) means that the number of BSs in VSN $\mathcal{C}_k$ has to be larger than u given Ω_k. Constraint (3.9) assures the remaining active BSs can guarantee the service of sleep BSs under certain BS-off strategy Ω_k, while the aggregated traffic loads of both active BSs and the covered sleep BSs cannot exceed overall predefined maximum traffic load l_{max} in (3.9), $l_{j,k}(t)$ denotes the instant traffic load in BS j of VSN $\mathcal{C}_k$ at time t.

3.3 Methodologies

In this chapter, we will introduce a solution to the problem (3.2) in a distributed manner, which is more feasible for practical implementations. The basic idea is to adaptively divide BSs into cell groups based on monitored traffic level.

Definition 2 (traffic level)**.** The traffic level is a term to describe the traffic condition in wireless networks based on the predefined measurements and the provided levels of service. We claim that the traffic condition can be translated to three levels with linguistic terms: low, moderate and high traffic levels.

Firstly, we describe the framework of the proposed solution, i.e., *reception*, *analysis*

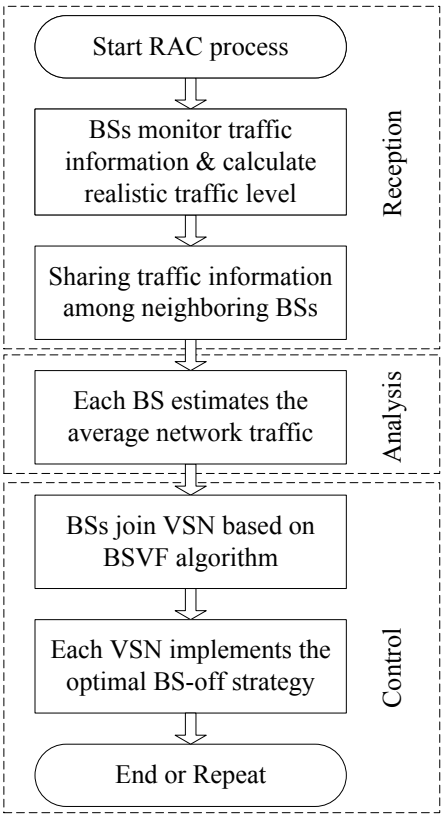


Figure 3.2: The strategy of distributed management of small cell networks, i.e., *reception, analysis* and *control* processes.

and *control* (RAC) process. Then, we provide the details in the calculation on BS realistic traffic level, estimation on mean network traffic, VSN formation procedure, BS-off strategy selection and BS on/off weight in following subsections.

3.3.1 Reception, Analysis and Control Process

The proposed solution framework consists of three procedures, called *reception*, *analysis*, and *control*.

- In *reception* stage, each BS collects information for VSNs formation, such as traffic load, VSN association status and the routing table from neighboring BSs. The gathered information is further calculated to obtain the realistic traffic levels in the corresponding BS surrounding area. The details on calculating the realistic traffic levels in a cell are shown in Section 3.3.2.

- In *analysis* stage, based on collected realistic traffic levels, each BS estimates the mean network traffic by using hidden Markov model (HMM) with details introduced in Section 3.3.3.

- In *control* stage, following the motivation to run separate BS-off strategies in each VSN, the entire network is partitioned into several VSNs. The VSNs' arrangement and the selected BS-off strategy in each VSN are updated with respect to varying traffic levels by adopting the proposed BSVF algorithm. All VSNs select and run the best BS-off strategy so as to switch off the underutilized BSs as many as possible to reduce the network power consumption.

Fig. 3.2 shows the flow chart of RAC process. Each VSN chooses the optimal BS-off strategy in an operating time slot. After that, with observed traffic variation, new VSNs will form and each VSN reselects its corresponding BS-off strategies. By doing these, the power consumption of the entire network can be reduced significantly.

3.3.2 Realistic Traffic Level Determination

In the traditional works considering cellular networks, the traffic level is measured merely by the number of MUs associated with each BS [50] [51]. Different from that,

in this chapter, the realistic traffic level of a cell is defined as a time-varying parameter by taking into consideration the number of MUs, the average distance between MUs and the BS, and the timeline. In reality, the association between a BS and MUs may become unstable when most MUs locate at the edge of a cell due to handover effects. Moreover, there is a significant difference by comparing the network traffic in daytime and night. Thus, if a cell has a high traffic level, it has to satisfy three conditions, i) the cell has a high traffic load, ii) most MUs are near the BS, and iii) the timeline lies in busy hours of a day. Since it is hard to directly determine the accurate bounds of the factors with respect to realistic traffic level of a cell, the calculation of the realistic traffic level is addressed by fuzzy inference system (FIS) as multiple inputs and one output problem [52] [53]. In this FIS, the number of MUs, the average distance between MUs and the BS, and the timeline are considered as fuzzy values [54]. Thus, the BS can make real-time decisions upon this soft computing method even with incomplete information due to, e.g., measurement errors.

For analysis purpose, we take a single BS into consideration and all other BSs follow the similar procedure. Let z_i be the input vector representing the information collected by the BS in *reception* stage with $z = (z_1, z_2, z_3)^{\Phi}$, where z_1 denotes the number of MUs, z_2 denotes the average distance between MUs and the BS, z_3 denotes the timeline and Φ is a linguistic label set. We initialize $\Phi(z_i), i = \{1, 2, 3\}$, with terms $\Phi(z_1) = \{light\,(QL), medium\,(QM), heavy\,(QH)\}$, $\Phi(z_2) = \{near\,(DN), middle\,(DM), far\,(DF)\}$ and $\Phi(z_3) = \{idle\,(TI), moderate\,(TM), busy\,(TB)\}$. Let l^{Φ} be the output vector, i.e. the realistic traffic level, and $\Phi(l) = \{very\,low\,(FVL), low\,(FL), medium\,(FM), high\,(FH), very\,high\,(FVH)\}$. We first determine the

degree of inputs to which they belong to each of the fuzzy sets $\mathcal{S}_i^{\Phi}$, and define it as

$$\mathcal{S}_i^{\Phi} = \left\{ \left(z_i, \psi_{\mathcal{S}_i^{\Phi}}(z_i) \right) | z_i \in \mathcal{D}_i \right\}. \tag{3.10}$$

In Eq. (3.10), the fuzzy set $\mathcal{S}_i^{\Phi}$ on a universe of discourse $\mathcal{D}_i$ is characterized by a set of ordered pairs, which are represented by a generic element z_i and its degree of membership function $\psi_{\mathcal{S}}(z_i)$ that takes values in the interval $[0, 1]$ with triangular and trapezoidal curves. Fig. 3.3 defines how each point in the input space is mapped to a membership value (or degree of membership) between 0 and 1. For z_1 and z_2, since there are precise values for the corresponding linguistic label, medium and middle, the triangular membership functions are applied. For the parameter timeline, the trapezoidal membership function is adopted because it is unlikely to define an exact timeline in a busy period. Since it is possible that some inputs have more than one corresponding value for the degree of the membership function, the *max* operator is adopted to decide the membership degree. The intersection of the degrees of membership functions within two fuzzy sets (e.g., $\mathcal{S}_1$ and $\mathcal{S}_2$) can be determined by [55]

$$\psi_{\mathcal{S}_1 \cap \mathcal{S}_2}(z) = \max\left(\psi_{\mathcal{S}_1}(z), \psi_{\mathcal{S}_2}(z)\right). \tag{3.11}$$

After the inputs are fuzzified, the *if-then* rules are required to map them to the cases of various term combinations. For instance, one of the *if-then* rules can be: if x_1 is $\mathcal{S}_1^{\Phi(z_1)}$, x_2 is $\mathcal{S}_2^{\Phi(z_2)}$and x_3 is $\mathcal{S}_3^{\Phi(z_3)}$, then l is Φ_l. In total, we have $3^3 = 27$ fuzzy rules for consideration. The fuzzy outputs with *if-then* rules are listed in Table 3.1.

The inputs to the defuzzification process are an aggregated fuzzy set. The Centroid

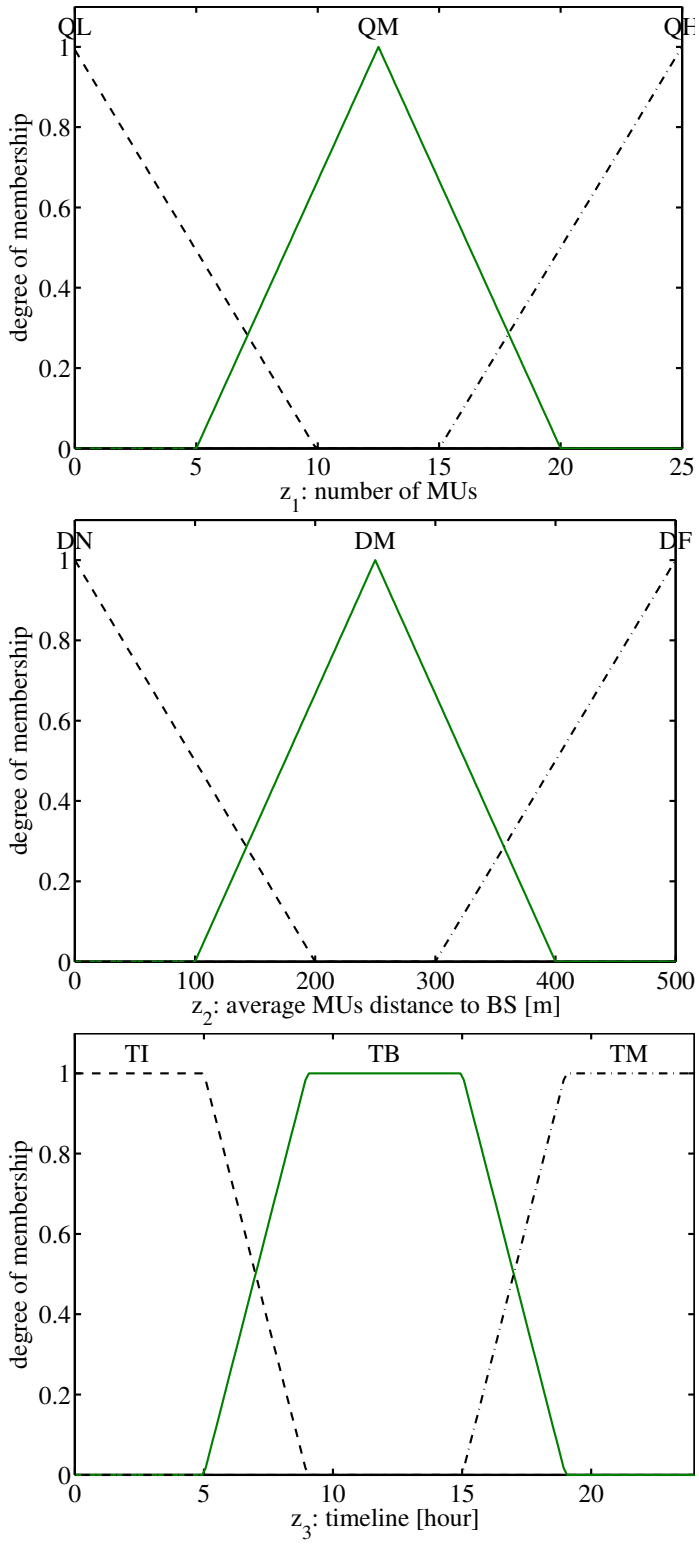


Figure 3.3: Membership functions of the input variables.

Table 3.1: The fuzzy rules to evaluate the realistic traffic levels based on the levels of the timeline including idle, moderate, and busy.

Timeline is idle			
	DN	DM	DF
QL	FL	FVL	FVL
QM	FM	FL	FVL
QH	FH	FM	FL

Timeline is moderate			
	DN	DM	DF
QL	FM	FL	FVL
QM	FH	FM	FL
QH	FVH	FH	FM

Timeline is busy			
	DN	DM	DF
QL	FH	FM	FL
QM	FVH	FH	FM
QH	FVH	FVH	FH

defuzzification method is adopted to determine the center of gravity given by

$$l = \frac{\sum_{i=1}^{I} z_i \psi_{\mathcal{S}_i^{\Phi}}(z_i)}{\sum_{i=1}^{I} \psi_{\mathcal{S}_i^{\Phi}}(z_i)}. \tag{3.12}$$

We then obtain a crisp output value l, which represents the realistic traffic level. Based on FIS, we can regulate mapping rules in terms of linguistic labels rather than numbers, and obtain the realistic traffic levels in cells by combining the key factors: time, user number and density. In Fig. 3.4, the defuzzified output of normalized realistic traffic level is presented by using Eq. (3.12). Note that even the traffic load in a cell is high, it may direct to different realistic traffic levels with respect to the other two factors.

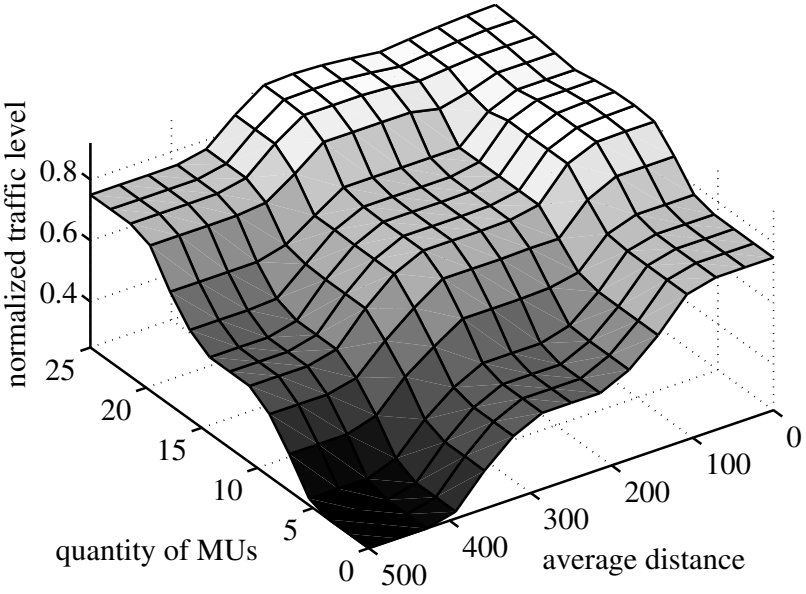


Figure 3.4: The mapping surface on the traffic levels of a cell.

3.3.3 Mean Network Traffic Estimation

The mean network traffic is defined as a time-varying parameter, denoting the average traffic level of the entire network in a time slot. The mean network traffic is a necessary condition for VSN formation such that the traffic level in any VSN can match the mean network traffic so as to select an optimal BS-off strategy network wide. However, the mean network traffic is unknown to individual BS in the distributed manner. The estimation can be viewed as a procedure that an individual BS needs to know the timely mean network traffic (i.e., the hidden states to the BS) under the knowledge of realistic traffic levels from the BS itself and all neighboring BSs (i.e., the observable states). Such viewing motivates the application of hidden Markov model (HMM) to estimate mean network traffic.

We introduce the following variables for modeling HMM. Let $t = \{0, 1, ..., T\}$ de-

note discrete time slots. Random variable $\mathcal{X} = (x_0, x_1, ..., x_T)$ denotes the hidden states representing the mean network traffic. $\mathcal{X}$ is stable during the estimation process. Random variable $\mathcal{Y}^{(q)} = (y_{q_0}, y_{q_1}, ..., y_{q_T})$ denotes a sequence of observed average traffic levels by a BS in observable state q_t. In other words, $y_{q_t} \in \mathcal{Y}^{(q)}$ indicates the average traffic level that a BS can calculate based on the realistic traffic levels of the neighboring BSs. The state transient matrix and the confusion matrix are denoted by $\mathcal{A} = (a_{nm}) = \mathcal{P}r(x_{n_t}|x_{m_{t-1}})$, and $\mathcal{B} = (b_{nm}) = \mathcal{P}r(y_n \mid x_m)$, respectively, and both $\mathcal{A}$ and $\mathcal{B}$ do not vary in time. Let initial state probability be π_n. Then, the modeled HMM can be represented as $(\pi, \mathcal{A}, \mathcal{B})$, which can be determined by empirical data [56].

Since we aim to find the most probable sequence of instant mean network traffic given an observation sequence, $\mathcal{X}$ is determined recursively by using Viterbi algorithm given $\mathcal{Y}^{(q)}$ and HMM with appropriate parameter matrices $(\pi, \mathcal{A}, \mathcal{B})$ [57]. In particular, each state at time t has a partial probability and a partial best path, which indicate the probability and best path of reaching a particular intermediate state in the trellis, respectively. The overall best path is obtained by choosing the state with the maximum partial probability and choosing its partial best path. We first calculate the partial probability of being in the state given $t = 0$ and the observable state q_0, i.e.,

$$\delta_0(n) = \pi(n) b_{nq_0}. \tag{3.13}$$

Then, for δ_t at time $t \geq 1$, the probability of the partial best path to a state n with the observable state q_t is calculated as

$$\delta_t(n) = \max_m \left(\delta_{t-1}(m) a_{nm} b_{nq_t}\right). \tag{3.14}$$

We also need to know in which state the system must have been at time $t-1$ if it is to arrive optimally at state n at time t. This recording is done by holding for each state a back pointer which points to the predecessor that optimally provokes the current state. The back pointer ϕ_t is formulated as

$$\phi_t(n) = \text{argmax}_m \left(\delta_{t-1}(m) a_{nm}\right). \tag{3.15}$$

Thus, we can determine which state at system completion ($t = T$) is the most probable as $x_t = \text{argmax}(\delta_T(x))$. For $t = \{T-1, T-2, ..., 0\}$, let $x_t = \phi_{t+1}(x_{t+1})$ be backtracking through the trellis following the most probable route. On completion, the sequence $x_0, x_1, ..., x_T$ will hold the most probable sequence of hidden states, i.e., the sequence of mean network traffic.

3.3.4 Self-organized VSN Forming Protocol

We consider the formation of VSNs by following two principles: *i*) BSs join a VSN if it makes the average traffic of the VSN approach the mean network traffic. After convergence, all the VSNs' average traffic will be balanced towards the mean network traffic; or *ii*) BSs join the VSNs so that the average traffic of VSNs lies in different traffic categories, e.g., high, middle and low. However, in the distributed manner, the challenge is that BSs cannot simply decide to group together without some measurements, e.g., the similarity measures. Thus, we proposed a BSVF algorithm based on firefly algorithm (FA). The FA is a heuristic optimization algorithm nature-inspired by the bioluminescence behavior of fireflies [58], in which the less bright firefly will move to the brighter one, and the brightness is an inverse ratio of the distance. Different from traditional FA, in the cellular network, BSs are deployed in hexagonal cell

layout and are standstill rather than randomly distributed and movable like fireflies. Also, due to the constraint that a BS can only share information with its neighboring BSs, if the VSN controller tries to communicate with its member BS, the communication distance is counted as the number of hops, h_{ij}, rather than Euclidean distance r in the classic FA. Therefore, we consider the brightness as an inverse ratio of the communication distance. In our proposed model, the derived BS realistic traffic level l and the estimated mean network traffic $\bar{l}$ are considered as the similarity measures.

Note that, we have defined all BSs form a grand coalition $\mathcal{N}$, and at the initial stage, there are K VSNs, $\{\mathcal{C}_k\}$, where $k = \{1, 2, ..., K\}$ and $K \geq 1$. Each VSN consists of c_k BSs. To facilitate the application of BS-off strategies and simplify the VSN forming process, K should be set as a small number so that c_k can be large enough to apply BS-off strategies. In practice, K can be determined based on the number of hot zones in the network. If there are total Z hot zones, the initial number of VSNs can be $K = Z + 1$. At initial stage, the VSN controller, $\mathcal{H}_k$, is selected randomly, and each BS selects the nearest $\mathcal{H}_k$ to join (in this case, header selection is similar to the traditional wireless sensor networks [59]).

Definition 3 (Fringe BS)**.** BS j in VSN $\mathcal{C}_k$ is a fringe BS, if at least one of its neighboring BS i belongs to a different VSN $\mathcal{C}_{k'}, k' \neq k$. BS j is confirmed to be a fringe BS if it satisfies the following criterion

$$\omega_{j,k}(t) + \sum_{i=1}^{n_j} \omega_{i,k}(t) \leq n_j, \forall j \in \mathcal{C}_k, \tag{3.16}$$

where n_j is the number of BS j's neighboring BSs.

Note that $\omega_{j,k} \in \{0, 1\}$ is a binary variable, which stands for the association between BS j and VSN $\mathcal{C}_k$. If $\omega_{j,k} = 1$, it means BS j is a member of VSN $\mathcal{C}_k$.

Otherwise, $\omega_{j,k} = 0$. Since each BS has only one associated VSN at any time, then,

$$\sum_{k=1}^{K} \omega_{j,k}(t) = 1, \forall j \in \mathcal{N}. \tag{3.17}$$

During the VSN formation period, only fringe BSs are allowed to join one of the neighboring VSNs. Such limitation can avoid isolated BSs or a BS is never separated from its associated VSN.

Algorithm 2 Pseudo code of BSVF algorithm

Input: $\mathcal{N}$,$\mathcal{T}$,$\mathcal{K}$,$\mathcal{C}_k$,$\mathcal{H}_k$
Output: $\mathcal{C}_k$,$\mathcal{H}_k$
Begin
while ($\tau <$ *MaxIteration*)
  for $i \in \mathcal{N}$ all BSs in $\mathcal{N}$
    if i is a fringe BS **then**
      for $j = 1$ to n_i (all n_i neighboring BSs of BS i)
        Find maximum attractiveness α_{ij}
        i join j's associate VSN, $\mathcal{C}_k^j \leftarrow \mathcal{C}_k^j \cup \{i\}$
        attractiveness α_{ij} changes with γ_{ij} and $\hat{h}_{ij}$
        update BS j's brightness β_j
      end for j
  end for i
  % *revise the location of VSN controller* $\mathcal{H}_k$
  for $k = 1$ to K (all K VSNs)
    for $i = 1$ to c_k (all c_k BSs)
      find maximum h_{ik}
      move $\mathcal{H}_k$ towards BS i
    end for i
  end for k
end while
evaluate average traffic level of VSNs
End

For two principles under consideration, brightness β can be defined respectively as i) the BS realistic traffic level of network, i.e., $\beta = l$, or ii) the absolute value of difference between BS realistic traffic level and the estimated mean network traffic, i.e., $\beta = |l - \bar{l}|$.

- **Case I**: $\beta = l$.

  In this case, the objective function for forming VSNs is formulated as

$$\min \mathcal{U}^1_{c_k \leq N} = |\mathcal{E}\left[\mathcal{L}(t)\right] - \bar{\mathcal{L}}_k(t)|, \forall k \in \mathcal{K}, \tag{3.18}$$

  where $\mathcal{E}\left[\mathcal{L}(t)\right]$ is the average network traffic. Note that it is different from the estimated mean network traffic $\bar{l}$ calculated by each individual BS. $\bar{\mathcal{L}}_k(t)$ is the average traffic of VSN $\mathcal{C}_k$. However, since $\mathcal{E}\left[\mathcal{L}(t)\right]$ is unknown in distributed network scenario, we reformulate the objective function as

$$\min \mathcal{U}^1_{c_{k1}, c_{k2} \leq N} = |\bar{\mathcal{L}}_{k1}(t) - \bar{\mathcal{L}}_{k2}(t)|, \forall k1, k2 \in \mathcal{K}. \tag{3.19}$$

  As the BS attractiveness is proportional to the intensity of brightness β observed by the neighboring BSs, the attractiveness function between BSs i and j, can be formulated as

$$\alpha^1_{ij} = \alpha_0 \exp\left(\frac{\gamma_{ij}}{\hat{h}_{ij}}\right), \tag{3.20}$$

  where γ_{ij} is defined as a traffic coefficient with respect to the practical average network traffic $\mathcal{E}\left[\mathcal{L}(t)\right]$ and brightness of BSs i and j. In this case, higher attractiveness is gained when the traffic level of compared BSs has a bigger gap. Thus, we define the traffic coefficient as $\gamma_{ij} = \frac{1}{\sqrt{\mathcal{E}[\mathcal{L}(t)]}}|\beta_i - \beta_j|$ so as to balance the traffic levels, i.e., the average traffic of formed VSN is approaching $\mathcal{E}\left[\mathcal{L}(t)\right]$. Since the average network traffic is not available in the distributed cellular network scenario, we consider a general average value for $\mathcal{L}$ for any BS, i.e., $\mathcal{E}\left[\mathcal{L}(t)\right] = \frac{1}{2}$ for normalized $\mathcal{L} \in [0, 1]$. If BS j is a member of VSN $\mathcal{C}_k$, then $\hat{h}_{ij}$ represents the number of hops from BS i to BS j's associated VSN

controller $\mathcal{H}_k$. Thus, we have $\hat{h}_{ij} = h_{jk} + h_{ij}$. α_0 is the attractiveness at $h = 0$, and usually we set $\alpha_0 = 1$ for simplicity. After comparison, BS i joins BS j with higher attractiveness. After that, the brightness of BS j from iteration τ to iteration $\tau + 1$ is updated as

$$\beta_j(\tau + 1) = \beta_j(\tau) + \frac{\beta_i(\tau)}{c_k + 1}. \tag{3.21}$$

When the convergence of VSN formation is completed, the average traffic in each VSN will be balanced.

Theorem 3.1 (VSNs with balancing average traffic)**.** *If the brightness is set to be the BS realistic traffic level, after running the BSVF algorithm, for any VSNs,* $k1$ *and* $k2$*, we have*

$$\lim_{n_{k1}, n_{k2} \to \infty} \Pr\left(|\bar{\mathcal{L}}_{k1} - \bar{\mathcal{L}}_{k2}| > \epsilon\right) = 0, \forall k1, k2 \in \mathcal{K}. \tag{3.22}$$

Proof. Let BSs realistic traffic levels $l_1, l_2, ..., l_{n_k}$ in any VSNs $\mathcal{C}_k$, $l_1, l_2, ..., l_{n_k}$ be a sequence of i.i.d random variables with average value $\bar{\mathcal{L}}_k = \frac{1}{n_k}(l_1 + l_2 + ... + l_{n_k})$. Based on Chebyshev's inequality on $\bar{\mathcal{L}}_k$, we have

$$\Pr\left(|\bar{\mathcal{L}}_k - \mu| \geq \epsilon\right) \leq \frac{\sigma^2}{n_k \epsilon^2}, \tag{3.23}$$

where μ is the common mean with $\mathcal{E}\left[\bar{\mathcal{L}}_k\right] = \mu$ and σ^2 is the variance of random variable sequence $l_1, l_2, ..., l_{n_k}$. From (3.23), we have

$$\Pr\left(|\bar{\mathcal{L}}_k - \mu| < \epsilon\right) = 1 - \Pr\left(|\bar{\mathcal{L}}_k - \mu| \geq \epsilon\right) \geq 1 - \frac{\sigma^2}{n_k \epsilon^2}. \tag{3.24}$$

As n_k goes to infinity, expression (3.24) approaches 1. Thus, we have $\Pr\left(|\bar{\mathcal{L}}_k - \mu|\right)$ $\to 0$ as $n_k \to \infty$. For any VSNs $k1$ and $k2$, we can obtain $\bar{\mathcal{L}}_{k1} \to \mu$ as $n_{k1} \to \infty$ and $\bar{\mathcal{L}}_{k2} \to \mu$ as $n_{k2} \to \infty$. Hence, expression (3.22) holds. □

- **Case II**: $\beta = |l - \bar{l}|$.

  In this case, the VSNs are formed based on different categories of traffic levels. The objective function can be formulated as

$$\max \mathcal{U}^2_{c_k \leq N} = |\mathcal{L}(t) - \bar{\mathcal{L}}_k(t)|, \forall k \in \mathcal{K}. \tag{3.25}$$

  The attractiveness function between BSs i and j can be formulated as

$$\alpha^2_{ij} = \alpha_0 \text{exp}\left(-\gamma_{ij}\hat{h}_{ij}\right). \tag{3.26}$$

  In this case, higher attractiveness is gained when two BSs have close traffic levels. The brightness update function is formulated similar to Eq. (3.21). When VSN formation procedure converges, some VSNs will gather the BSs with high traffic levels, while others collect the BSs with low or moderate traffic levels.

Theorem 3.2 (VSN with heterogeneous average traffic)**.** *When the brightness is set to be the absolute value of the difference between BS realistic traffic level and estimated mean network traffic, for any VSNs $k1$ and $k2$, if $l_i > \bar{l}_i$, $l_i \in (l_1, l_2, ..., l_{n_{k1}})$ in VSN $\mathcal{C}_{k1}$, and $l_j < \bar{l}_j$, $l_j \in (l_1, l_2, ..., l_{n_{k2}})$ in VSN $\mathcal{C}_{k2}$, we have*

$$\lim_{n_{k1}, n_{k2} \to \infty} \text{Pr}\left(\bar{\mathcal{L}}_{k1} - \bar{\mathcal{L}}_{k2}\right) > 0, \ \forall k1, k2 \in \mathcal{K}. \tag{3.27}$$

Proof. It is obvious that the BSs with the realistic traffic levels under or over the BSs estimated mean network traffic are separately grouped in terms of the brightness and attractiveness functions in case II. It is the necessary and sufficient conditions for the assumption that $l_i > \bar{l}_i$, $l_i \in (l_1, l_2, ..., l_{n_{k1}})$ in VSN $\mathcal{C}_{k1}$, and $l_j < \bar{l}_j$, $l_j \in (l_1, l_2, ..., l_{n_{k2}})$

in VSN $\mathcal{C}_{k2}$. Since $\mathcal{E}[\mathcal{L}] \approx \frac{1}{n_{k1}}(\bar{l}_1 + \bar{l}_2 + ... + \bar{l}_{n_{k1}}) \approx \frac{1}{n_{k2}}(\bar{l}_1 + \bar{l}_2 + ... + \bar{l}_{n_{k2}})$ as $n_{k1}, n_{k2} \to \infty$ we have $\bar{\mathcal{L}}_{k1} = \frac{1}{n_{k1}}(l_1 + l_2 + ... + l_{n_{k1}}) > \mathcal{E}[\mathcal{L}]$ as $n_{k1} \to \infty$ and $\bar{\mathcal{L}}_{k2} = \frac{1}{n_{k2}}(l_1 + l_2 + ... + l_{n_{k2}}) < \mathcal{E}[\mathcal{L}]$ as $n_{k2} \to \infty$. This completes the proof. □

Corollary (3.3.4)**.** *The formation of virtual small network (VSNs) leads to better power saving results than the case without VSNs under the BS-off schemes.*

The **Corollary** 3.3.4 can be proofed by an example. Let a cellular network be partitioned into three VSNs, $\{\mathcal{C}_{k1}, \mathcal{C}_{k2}, \mathcal{C}_{k3}\}$. The average traffic of the VSNs is set to be very different with $\bar{\mathcal{L}}_{k1} > \bar{\mathcal{L}}_{k2} > \bar{\mathcal{L}}_{k3}$. According to the implementation of BS (v, u)-off strategy, we assume the best BS-off schemes for each VSN are (v, u)-off, $(v+1, u)$-off and $(v+2, u)$-off, respectively, and $u > v+2$. The network power saving ratio can be simply derived as $\frac{1}{3}(\frac{v}{u} + \frac{v+1}{u} + \frac{v+2}{u}) = \frac{v+1}{u}$. In fact, this scenario is similar to the case II, which groups the BSs into different traffic categories. In case I, after the convergence, all the VSNs have similar average traffic level, which is approximately to $\bar{\mathcal{L}}_{k2}$. Thus, with the corresponding best BS-off scheme $(v+1, u)$-off, the network power saving ratio is approximately $\frac{v+1}{u}$. However, in the traditional network without VSNs, the unified BS-off scheme is adopted. In order to satisfy the entire network demand, the best BS-off scheme is obviously (v, u)-off with the network power saving ratio of $\frac{v}{u}$.

At the end of each iteration, the location of VSN controller $\mathcal{H}_k$ should be updated to guarantee that the VSN controller always stays in the geological center of the VSN. $\mathcal{H}_k$ needs to maintain a link table which keeps a record of hopping topologies for the member BSs. Based on the link table, $\mathcal{H}_k$ will move towards BS i with maximum h_{ik}. The detailed process of BSVF algorithm is presented in Algorithm 2.

3.3.5 The Modified BS-off Matching Scheme

After VSN formation process, each VSN controller selects the best BS-off strategy for its VSN. The selection can be formulated as an optimization problem, which aims to minimize total system power consumption at time slot t, $\mathcal{P}_{nw}^{t}$. We can reformulate Eq. (3.2) by focusing on the selection of BS-off scheme[1].

$$\min \mathcal{P}_{nw}^{t} = \sum_{k=1}^{K} \sum_{j=1}^{c_k} \left((1+\rho)\mathcal{P}_{sta}^{(\Omega_k)} + \left[\mathcal{P}_{dyn}^{(\Omega_k)}(t) \right]_{(j,k)} \right) \tag{3.28}$$

$$\sum_{k=1}^{K} \sum_{j=1}^{c_k} \left((1+\rho)\mathcal{P}_{sta}^{(\Omega_k)} + \eta \cdot \omega_{j,k} \cdot \lambda_{j,k} \cdot l_{j,k}^{(\Omega_k)}(t) \right) \tag{3.29}$$

$$\text{s.t.} \quad \sum_{j=1}^{N} \omega_{j,k} \geqslant 2, \forall k = 1, 2, \ldots, K. \tag{3.30}$$

$$(3.3) \sim (3.9),$$

where η stands for the coefficient of BS dynamic power consumption $\mathcal{P}_{dyn}$. $\lambda_{j,k} \in \{0, 1\}$ indicates BS on/off state, in which $\lambda_{j,k} = 1$ means a BS j in VSN $\mathcal{C}_k$ is active, otherwise $\lambda_{j,k} = 0$. Constraint (3.30) means that any VSN must have at least 2 associated BSs so that the minimum BS-off matching scheme, i.e., (1,2)-off can be operated. The other constraints follow a similar discussion in Eq. (3.2). In practice, since the set of feasible BS-off strategies Ω is finite, exclusive searching can be adopted in all potential candidates to find the best one for implementation.

[1]The modified BS-off matching scheme may follow the similar discussion in Chapter 2. However, since the system model, definitions and variables have changed, in this chapter, the detailed formulation and explanation are rewritten accordingly.

3.3.6 BS Power On/off Weight

In the BS-off matching scheme, the number of switched-off BSs can be determined in each VSN. However, for a specific BS, whether it is power on or off is not carefully decided yet. Therefore, we need to further determine which BS is better to be switched off. Note that too often transitions between the active and sleep modes may lead to more power consumption when considering the power cost in the BS transient period. Also, frequent transitions of BS working modes may affect system stability. Thus, we introduce the concept of BS on/off weight, which can be simply integrated to the BSVF algorithm. The BS on/off weight, denoted by $\mathcal{W}_{on}$, is defined as the probability that a BS turns to or stays in active mode. n_{con} is a constant for adaptive control of $\mathcal{W}_{on}$. n_{con} can be either an incremental or decremental value for the BS on/off weight learning procedure. The design principles to prevent a BS in frequent working mode transitions are summarized as *i*) if a BS is in active mode currently, its probability of staying active in next operation period is increasing; *ii*) if a BS is in sleep mode, its probability of switching power on in next operation period is decreasing. Such principles can make a BS enlarge the probability of keeping its working modes unchanged during each operation period.

At initial stage, each BS is assigned to a random value of $\mathcal{W}_{on} \in (0, 1)$. The value of n_{con} can be set as a very small value. After that, BSs will update the value of $\mathcal{W}_{on}$ as a learning procedure at the end of each operation stage, which is required to obey the following rules.

- If $0 < \mathcal{W}_{on} < 1$, $\mathcal{W}_{on} = \mathcal{W}_{on} + n_{con}$ when a BS is in active mode or $\mathcal{W}_{on} = \mathcal{W}_{on} - n_{con}$ when a BS is in sleep mode.

- If $\mathcal{W}_{on} \leq 0$, $\mathcal{W}_{on} = \max(0,\ W_{on} + n_{con})$.

- If $\mathcal{W}_{on} \geq 1$, $\mathcal{W}_{on} = \min(1,\ W_{on} - n_{con})$.

Thus, BSs with a larger value of $\mathcal{W}_{on}$ possess a high probability to be switched on or kept active mode. On the other hand, a BS is more likely to be powered off or keeping in sleep mode with small $\mathcal{W}_{on}$. Note that BS as a VSN controller should be kept in active mode during an operation period with $\mathcal{W}_{on} = 1$.

3.4 Simulation Results

In this section, simulation results are demonstrated to evaluate the performance of the proposed RAC process integrating BSVF algorithm, BS-off matching and BS power on/off weight schemes in wireless cellular networks. The simulation scenario is based on the evaluation methodology described in [60], while the proposed distributed BS management strategy is compared with the traditional unified BS-off strategy, called unified BS-off [16] [17].

The micro hexagonal cell layout deployed in an urban environment is considered. Followed by the power consumption characteristic of a typical micro BS, its cell radius is set to $\mathcal{R} = 500$ m, which is corresponding to $\mathcal{P}_{sta} = 237$ watt for a typical micro BS. The power transient cost ratio is set to be $\rho = 0.05$. The path loss model is compliant with micro-cell test environment in ITU report with a center frequency of 2.655 GHz [60] and the receiver sensitivity of MUs is -120 dB. The maximum number of MUs can be served by a BS is $l_{max} = 25$, which is corresponding to assigning 5 MHz bandwidth in an LTE system. The minimum number of hot spots is set to $Z = 4$.

The arrival process of MUs follows Poisson distribution.

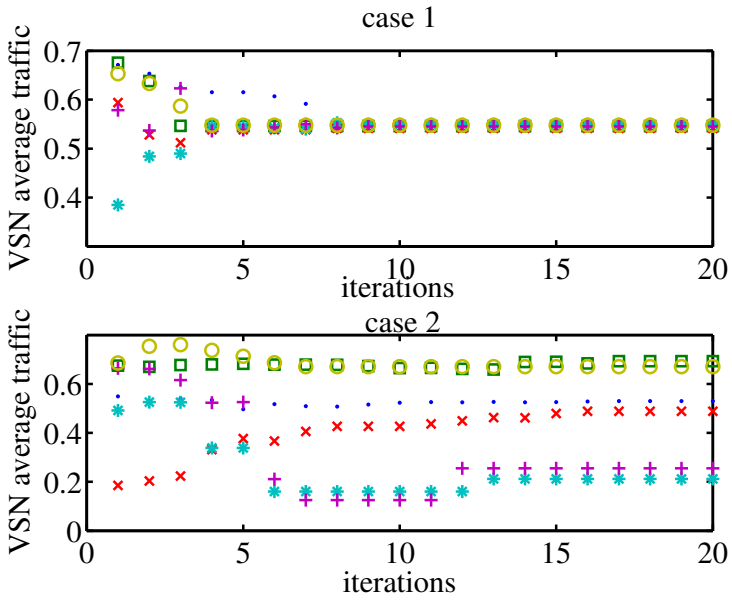


Figure 3.5: Convergence rates on VSNs' average traffic in two cases of the BSVF algorithm, with $N = 300$, $K = 6$. Different marks show the average traffic of VSNs.

Fig. 3.5 shows the convergence on average traffic levels of VSNs when the BSVF algorithm is running under two cases with different brightnesses. Note that in case I, the average traffic of all VSNs approach a similar point as the iteration proceed, i.e., all VSNs' average traffic levels tend to the mean network traffic, which satisfies **Theory** 3.22. That means, in case I, the VSNs may choose the same BS-off strategy due to similar average traffic. However, in case II, the average traffic levels of VSNs are obviously variable, which indicates that the VSNs form in different traffic categories, i.e., low, moderate and high traffic levels. In case II, each VSN may select the most aggressive BS-off strategy to match individual traffic level. Another observation for both cases is the time spent on VSN formation is quite small because commonly the convergence of each VSN can be completed under 10 iterations.

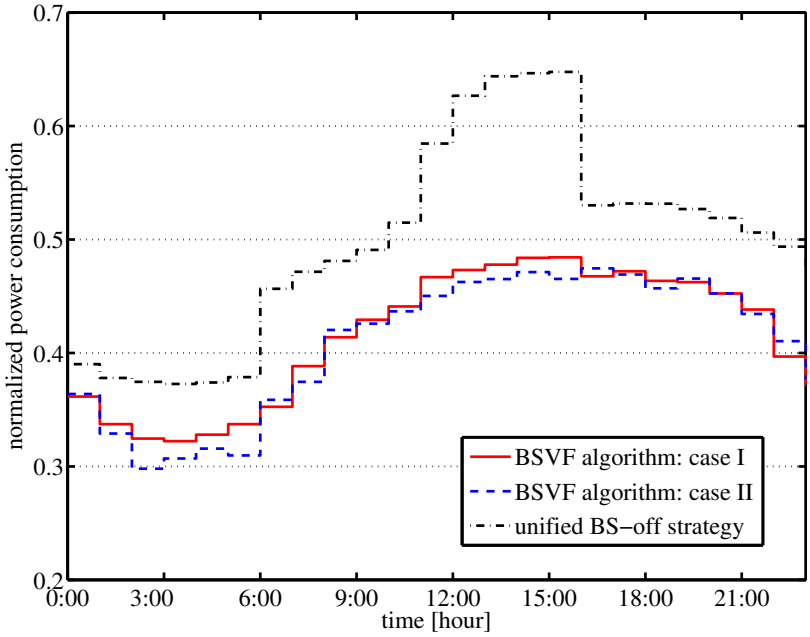


Figure 3.6: Comparison of power consumptions in case I and II with the unified BS-off strategy, $N = 200$, $K = 5$.

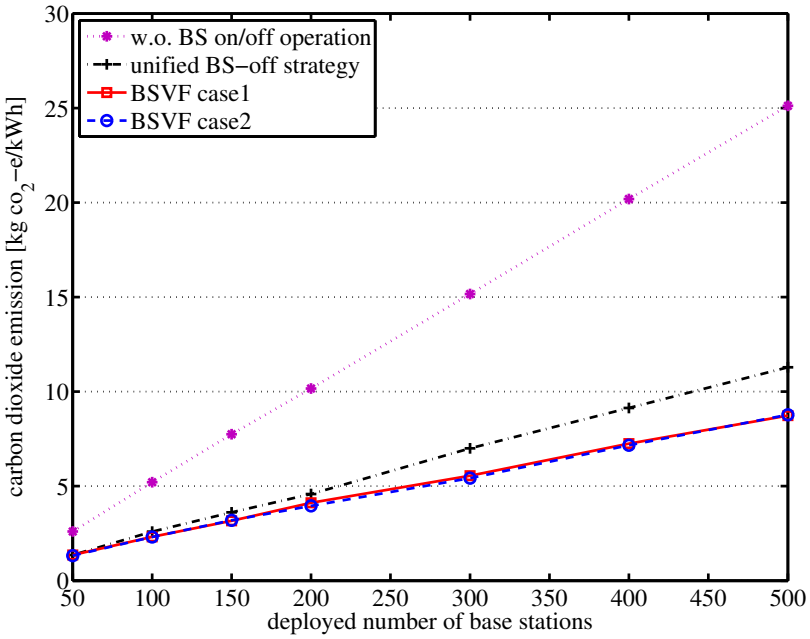


Figure 3.7: The CO_2-e changes with increasing number of the deployed micro-BSs, $K = 5$.

Fig. 3.6 compares the power consumption of the proposed BSVF algorithm and traditional unified BS-off strategy. Simulation results are normalized with respect to the network power consumption without any VSN formation and BS-off mechanisms. Compared to the unified BS-off strategy, it is apparent that the total consumed power of the network can be significantly reduced by the proposed protocol, particularly when the network lies in extremely high and low traffic level, which may relate to the office hours and nighttime, respectively. We also observe that case II is slightly superior to case I. It is because case II can better fit for the traffic fluctuations especially when the traffic in the whole network is more unevenly distributed. In this scenario, it is easier to form VSNs in different traffic categories than to balance the traffic levels of all VSNs. In short, the proposed protocol can save 60% power consumption compared to that without any VSN formation and BS-off mechanisms, and both proposed cases outperform traditional unified BS-off strategy with 10% power saving enhancement.

We further show CO_2-e in daily statistics with different number of BSs in Fig. 3.7. CO_2-e equals 0.457 kg CO_2 per kWh based on the report on electricity emission factor [61] [62][2]. As the number of BSs increases, the emitted CO_2 of the proposed protocol is obviously less than the unified BS-off strategy and far less than the case without BS on/off operation. When there exist $N = 500$ BSs in the network, the proposed protocol can reduce the emission of 16.5 kg CO_2 daily.

In Fig. 3.8, we present the power saving ratio with the choice of initializing a different number of VSNs, e.g., $N = 200$. Our assumption is proved that if we

[2]Based on global electricity production statistics, 500 g CO_2-e per kWh is emitted in average. However, in reality, the CO_2-e emission per kWh vary depending on the country or region where the electricity is produced.

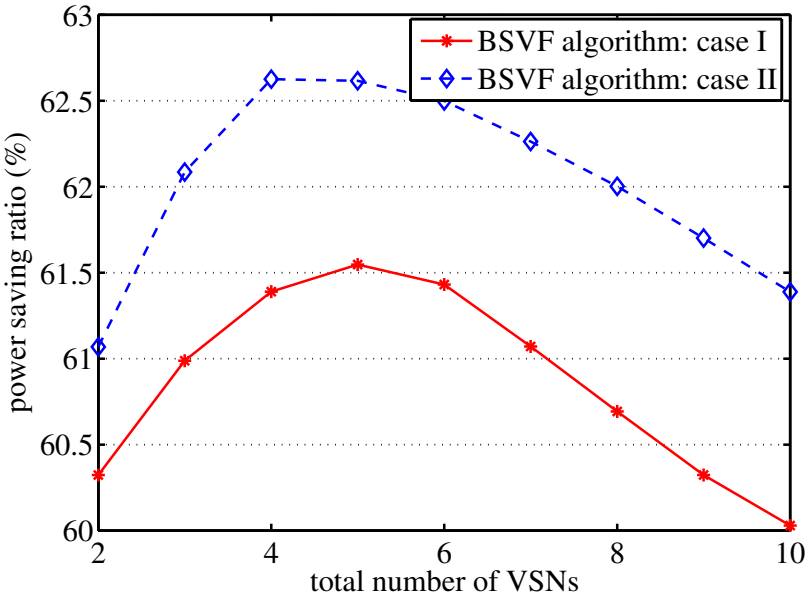


Figure 3.8: The varied tendency of power saving ratio in case I and II follows an increasing number of VSNs, $N = 200$.

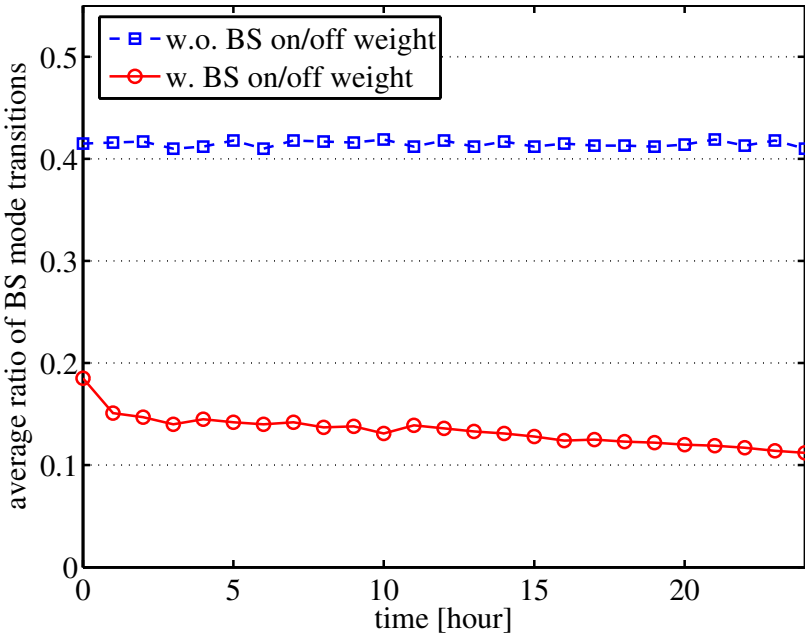


Figure 3.9: Average ratio of BS active and sleep modes' transitions with/without BS on/off weight, $N = 300$, $K = 5$.

have total Z hot spots in the network, the optimal number of partitioned VSNs is $K = Z + 1$. Less number of VSNs is not able to adaptively cover the heterogeneous traffic distribution, while too many VSNs will cause performance degradation due to more aggregated power cost in BS transient period.

Fig. 3.9 shows the average ratio of BSs mode transitions between the active and sleep modes following the operation time slot with $n_{con} = 0.01$. Note that at each operation period, average 42% BSs have to change between active and sleep modes without BS on/off weight. However, by taking into consideration the BS on/off weight, the average BS modes transition rate is reduced significantly at 13.5%. Such results validate that the BS can avoid frequently shifting between the active and sleep modes with the BS on/off weight learning procedure.

Chapter 4

Strategy of Adaptive Service Rate and Vacation Length for Energy-efficient HeNB based on Queueing Analysis

In this chapter, to enhance the energy-efficiency in the femtocells, we propose the strategy of adaptive service rate and vacation length based on queueing analysis. First, the HeNB on/off mechanism is analyzed by formulating an MAP/PH/1/K queueing model. The formulated discrete time Markov chain models the hybrid access and transmitter power on/off mechanism in order to understand how the service of priority users can affect the HeNB on/off transitions. Then, we evaluate some key performance metrics such as user waiting time, blocking probability and system busy cycle. After that, we propose an adaptive service rate and vacation length

(ASV) method, which can improve the HeNB energy efficiency while guaranteeing QoS provisioning. Table 4.1 summarizes the important notations in this chapter for ease of reference.

4.1 System Model

Consider a macrocell-based cellular network with an overlaid HeNB deployment for covering hotspots or indoor areas. The HeNB enables hybrid access mode and transmit power on/off mechanism [35, 63]. The hybrid access mode consists of two user groups, i.e., high priority users (HP users) from the CSG and low priority users (LP users) from open access. These two user groups are different in terms of arrival rate, service rate and priority of service. Define $\mathbb{R} = \{r_i\}$, $i = 0, 1, ..., \mathcal{N}$ and $\mathcal{N} < \infty$, as a finite set of available service rates at the HeNB in ascending order, and $\mathbb{P} = \{p_i\}$, $i = 0, 1, ..., \mathcal{N}$, as the corresponding consumed power when the service rate is r_i. In the transmit power on/off mechanism, there are two HeNB operation modes, i.e., transmitter power on (service mode) and off (vacation mode)[1]. At the beginning of each vacation period, the HeNB will choose a suitable vacation length from a finite set of available vacation lengths, $\mathbb{V} = \{v_i\}$, $i = 1, 2, ..., \mathcal{M}$ and $\mathcal{M} < \infty$. At the end of each vacation period, there is a listening period to monitor the buffer occupancy to determine whether the vacation terminates or not. The powers consumed in vacation and listening periods are denoted by p_v and p_l, respectively. We consider a discrete time system and adopt the time division duplex (TDD) scheme to take advantage of

[1]Describing the HeNB non-operational mode has multiple terms in the literature, e.g., power off mode, dormant or sleep mode. In this chapter, we adopt the term 'vacation mode' to better match the formulation of queueing model.

3GPP LTE technology [14]. Each slot in time is long enough to complete service of one typical user.

For analysis purpose, assume that user arrivals can only occur at discrete time epochs $t = n^-, n = 0, 1, 2, \ldots$, and the service starting and ending times can only occur at discrete time epochs $t = n^+, n = 0, 1, 2, \ldots$. Considering a causal scheduler, the service to a new arrival will start from the next slot. Such model is classified as a late arrival system with delayed access (LAS-DA) [42]. In one time slot, the HeNB may have no user arrival, just one user arrival of either type, or two user arrivals with one of each type. We model the interarrivals of users as Markovian arrival process (MAP) in order to capture the potential correlation among arrivals. The modeled MAP can be represented by four matrices $\mathcal{D}_{0,0}$, $\mathcal{D}_{1,0}$, $\mathcal{D}_{0,1}$, and $\mathcal{D}_{1,1}$ [43], all with dimensions $n_a \times n_a$, where n_a denotes the order of probability matrix of MAP. $\mathcal{D}_{\text{HP,LP}}$ with binary variables HP and LP denotes the state of user arrivals. HP(LP) = 1 represents an HP(LP) user arrival. Otherwise, HP(LP) = 0. Let $\mathcal{D} = \mathcal{D}_{0,0} + \mathcal{D}_{1,0} + \mathcal{D}_{0,1} + \mathcal{D}_{1,1}$, and $\mathcal{D}_0 = \mathcal{D}_{0,0}$. Note that D is stochastic. The probabilities of the HP and LP user arrivals in a slot are $\lambda_1 = \lambda(\mathcal{D}_{1,0} + \mathcal{D}_{1,1})\boldsymbol{e}$ and $\lambda_2 = \lambda(\mathcal{D}_{0,1} + \mathcal{D}_{1,1})\boldsymbol{e}$, respectively, where λ is the solution of $\lambda = \lambda\mathcal{D}$ and $\lambda\boldsymbol{e} = \boldsymbol{e}$. Here, $\boldsymbol{e}$ denotes a column vector of ones with an appropriate dimension.

The service of both HP and LP users is assumed to follow a phase-type (PH) distribution, denoted by $(\beta_m, \mathcal{S}_m)$ of order n_s, where $m = 1, 2$ denote the HP and LP users, respectively. β is the initial probability vector of order $1 \times n_s$ satisfying $\beta_m\boldsymbol{e} = 1$, and $\mathcal{S}$ of order $n_s \times n_s$ denotes the incomplete service. The service completion is denoted by $\mathbf{s}_m$ of order $n_s \times 1$ satisfying $\mathbf{s}_m = \boldsymbol{e} - \mathcal{S}_m\boldsymbol{e}$. The average service rate for

type m users is $\mu_m^{-1} = \beta_m(\mathcal{I} - \mathcal{S}_m)^{-1}\boldsymbol{e}$, where $\mathcal{I}$ denotes the identity matrix.

Each HeNB power off period may consist of multiple vacations. Each vacation follows a PH distribution denoted by $(\delta, \mathcal{V})$ of order n_v. The mean vacation length is $v^{-1} = \delta(\mathcal{I} - \mathcal{V})^{-1}\boldsymbol{e}$. In addition, to better match the differential user priorities in the HeNB hybrid access mode, it is allowed that an HP user arrival can interrupt the server vacations, called 'vacation termination' policy. We use stochastic matrix $\mathcal{Q}$ of dimension $n_v \times 1$ to depict the vacation termination process.

The HeNB maintains a buffer of size k, $k < \infty$, for both HP and LP users. The arriving users will join the queue if the buffer has the vacancy or be blocked otherwise. A complete buffer sharing policy is adopted, which means there is no buffer reservation for any type of users. The users in the same type are served according to the arrival order, i.e., first-come first-serve (FCFS). A non-preemptive discipline is adopted such that no service for an LP user is going to be started if there is an HP user in the system. However, if the service of an LP user has started, the current service of the LP user cannot be interrupted [44]. In addition, the queueing model of the HeNB implements exhaustive service and ungated vacation, which needs to satisfy the following disciplines.

- *Exhaustive service*: all users should be served including those waiting in the queue and those arrive after the server returns from vacation.
- *Ungated and multiple vacations*: the HeNB cannot take any vacation until the buffer becomes empty. When the buffer is still empty after a vacation, the HeNB starts another vacation immediately.
- *LP user arrival during vacation*: the HeNB has no obligation to terminate a

vacation and start to serve the LP users that arrive during the vacation period. Nevertheless, the newly arrived LP users should be buffered in the queue.

Table 4.1: Summary of key notations

Notation	**Physical meaning**
t	Discrete time epoch
k	HeNB buffer size
HP, LP	High priority, low priority users
$m = 1, 2$	Indicating HP, LP users as footnotes
λ_1, λ_2	HeNB arrival rate for HP/LP users
μ_1, μ_2	HeNB service rate for HP/LP users
$\boldsymbol{e}$, $\mathcal{I}$	A column vector of ones, identity matrix
$\mathcal{D}_{0,0}$, $\mathcal{D}_{1,0}$, $\mathcal{D}_{0,1}$, $\mathcal{D}_{1,1}$	Markovian arrival process (MAP) for HP/LP users
$(\beta_m, \mathcal{S}_m)$	Phase-type (PH) distribution for HeNB service
$(\delta, \mathcal{V})$	PH distribution for HeNB vacation
$\mathbf{s} = \boldsymbol{e} - \mathcal{S}\boldsymbol{e}$	Service completion in PH distribution
$\mathbf{v} = \boldsymbol{e} - \mathcal{V}\boldsymbol{e}$	Vacation completion in PH distribution
$\mathcal{Q}$	A stochastic matrix denotes vacation termination
n_a, n_s, n_v	Dimensions of matrices for arrival, service and vacation
Δ	System state space
l_m	Number of type-m users in the system
l	Total number of users in the system
ρ	Traffic intensity
z, $\theta \in \{s_m, v\}$	Phase of arrival, phase of service or vacation

$\mathcal{P}$	State transition probability matrix
$\mathcal{A}$, $\mathcal{B}$, $\mathcal{C}$	Sub-matrices of transition matrix $\mathcal{P}$
$\mathcal{L}_1$, $\mathcal{L}_2$	Probability of HP/LP users in the system
$\overline{\mathcal{L}}_1$, $\overline{\mathcal{L}}_2$	Average number of HP/LP users in the system
$\overline{\mathcal{W}}_1$, $\overline{\mathcal{W}}_2$	Average waiting time of HP/LP users in the system
$\mathcal{P}_{B_1}$, $\mathcal{P}_{B_2}$	Blocking probability of HP/LP users
E_{sys}	System energy efficiency
$\mathcal{E}_n^b$	Energy efficiency of busy cycle n
$\bar{t}_b$, $\bar{t}_s$, $\bar{t}_v$	Mean time of system busy cycle, service and vacation
p_v, p_l	Power consumption of vacation and listening periods
$\mathbb{R} = \{r_i\}$	Finite set of service rates in ascending order
$\mathbb{P} = \{p_i\}$	Finite set of consumed powers matching service rates
$\mathbb{V} = \{v_i\}$	Finite set of available vacation lengths

4.2 Queueing Analysis

In this section, we formulate the queueing model and derive the key performance measures including queue length, user waiting time, blocking probability and system busy cycle.

4.2.1 Queueing Model Formulation

The aforementioned system paradigm can be modeled as an MAP/PH/1/k queue with multiple vacations and non-preemptive priorities. Let $l_m \in \{0, 1, \ldots, k\}$, $m =$

$1, 2$, indicate the number of type-m users in the system, and $l_1 + l_2 \leqslant k$. Let z denote the arrival phase of both types of users, and $\theta = \{s_m, v\}$ denote the phase of server working mode, where s_m represents the phase of service for type-m users and v represents the phase of vacation. Note that since the server is in either service or vacation mode, only one of s and v can be referred at any time instant. Then, the behavior of the system can be described by a four-dimensional stochastic process as $\Delta = (l_1, l_2, z, \theta)$ with $\{0 \leqslant l_1 \leqslant k, 0 \leqslant l_2 \leqslant k - l_1\}$.

We can classify the system state space Δ into four categories: *i*) the system is empty and on vacation, *ii*) the system is in service and only the LP users exist in the system, *iii*) the system is in vacation with the LP users in the system, and *iv*) there are HP users in the system. In summary, the state space can be rewritten as

$$\Delta = \Delta_0^v \cup \Delta_1^v \cup \Delta_1 \cup \Delta_2 \tag{4.1}$$

where $\Delta_0^v = (0, 0, z, v)$ means an empty system with arrival in phase z and the server is on vacation in phase v. $\Delta_1^v = (0, l_2, z, v)$, $l_2 \in [1, k]$, indicates that the server is on vacation of phase v with only the LP users waiting in the system. $\Delta_1 = (0, l_2, z, s_2)$, $l_2 \in [1, k]$, means there is no HP user and at least one LP user exists in the system with service in phase s_2. $\Delta_2 = (l_1, l_2, z, s_m)$, $l_1 \in [1, k]$, $l_2 \in [0, k - l_1]$, represents there are at least one HP user and l_2 LP users in the system with one user of type-m in service and the server is in phase s_m. Note that Δ_2 also includes the case when there are only HP users in the system with $l_2 = 0$ and $m = 1$. Note that according to the vacation termination policy, if an HP user joins in the queue when the system is on vacation, the server has to exit current vacation mode and the system state will transit from Δ_1^v to Δ_2.

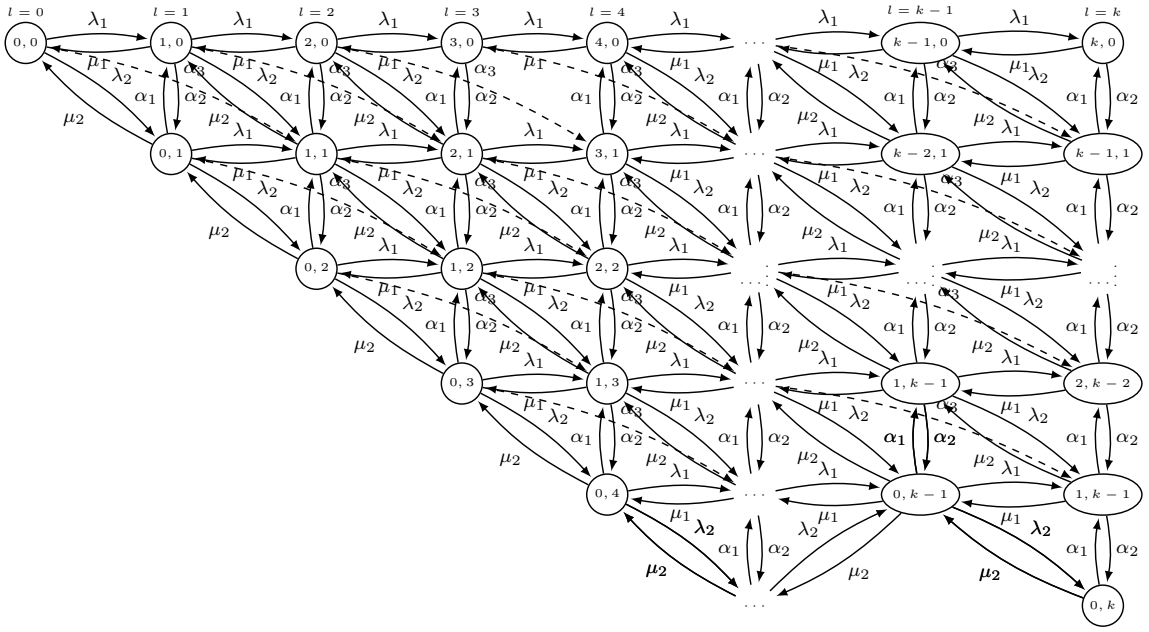


Figure 4.1: State transitions of MAP/PH/1/k queue with user priorities, where the numbers in circles represent (l_1, l_2), $l = l_1 + l_2$, $\alpha_1 = \lambda_1 + \mu_2$, $\alpha_2 = \lambda_2 + \mu_1$ and $\alpha_3 = \lambda_1 + \lambda_2$. Particularly, the state transition of α_3 is presented in dash line, and the transitions between any node and itself are omitted for clarity.

The state transitions of the constructed MAP/PH/1/k queue with user priorities are shown in Fig. 4.1, where the numbers in circles represent (l_1, l_2), and $l = l_1 + l_2$ indicates the total number of users in the system. The state transition probability matrix $\mathcal{P}$ of the modeled MAP/PH/1/k queue can be described as

$$\mathcal{P} = \begin{bmatrix} \mathcal{A}_0 & \mathcal{B}_0 & & & & & \\ \mathcal{C}_0 & \mathcal{A}_1 & \mathcal{B}_1 & & & & \\ & \mathcal{C}_1 & \mathcal{A}_2 & \mathcal{B}_2 & & & \\ & & \ddots & \ddots & \ddots & & \\ & & & \mathcal{C}_{k-2} & \mathcal{A}_{k-1} & \mathcal{B}_{k-1} \\ & & & & \mathcal{C}_{k-1} & \mathcal{A}_k \end{bmatrix} \tag{4.2}$$

where $\mathcal{A}_n$, $n = 0, 1, \ldots, k$, $\mathcal{B}_n$, $n = 1, 2, \ldots, k-1$, and $\mathcal{C}_n$, $n = 0, 1, \ldots, k-1$, are

submatrices. $\mathcal{A}_n$ denotes the transitions under the condition that there are $l_1 = n$ HP users in the system, $\mathcal{B}_n$ denotes the transitions when an HP user arrives (i.e., $l_1 = n \rightarrow l_1 = n+1$) in the system and there is no service completion, and $\mathcal{C}_n$ denotes the transitions when the service of an HP user is completed (i.e., $l_1 = n \rightarrow l_1 = n-1$). All blank areas are zeros and only $\mathcal{A}_n$ are square matrices. $\mathcal{A}_n$, $\mathcal{B}_n$, and $\mathcal{C}_n$ involves the states in $\Delta_0^v \cup \Delta_1 \cup \Delta_1^v$, Δ_2, and $\Delta_0^v \cup \Delta_1 \cup \Delta_2$, respectively.

We use submatrix $\mathcal{A}_0$ as an example to show the derivation of the transition matrix. Note that $\mathcal{A}_0$ denotes the probabilities that system states transform from $(0, l_2)$ to $(0, l_2')$ when there are no HP users involved in state transitions. The derivation of $\mathcal{A}_0$ consists of following seven cases.

- *Case I.* Empty system with the server on vacation, i.e., $(l_1, l_2) = (0, 0)$

  In case I, there is no arrival from either class of users (the event probability is $\mathcal{D}_0$). The server may be in the middle of vacation (with probability $\mathcal{V}$), or at the end of one vacation so that the system starts another vacation due to empty buffer (with probability $\mathbf{v}\delta$). Thus, the transition matrix in case I can be written as

  $$\mathcal{A}_0^{00} = \left[\begin{array}{cc} 0, & \mathcal{D}_0 \otimes (\mathbf{v}\delta + \mathcal{V}) \end{array} \right] \tag{4.3}$$

  where $\otimes$ denotes Kronecker product, and 0 denotes that the probability of the server in service is 0.

- *Case II.* State transition from $(l_1, l_2) = (0, 0)$ to $(0, 1)$

  In case II, a new LP user arrives with a probability $\mathcal{D}_{0,1}$. Due to the low priority of the new arrival, the server may keep staying in the vacation (with probability

$\mathcal{V}$) or finish the current vacation and be ready to serve the LP user in the next slot (with probability $\mathbf{v}\beta_2$). Thus, the transition matrix in case II is

$$\mathcal{A}_0^{01} = \begin{bmatrix} \mathcal{D}_{0,1} \otimes (\mathbf{v}\beta_2) & \mathcal{D}_{0,1} \otimes \mathcal{V} \end{bmatrix}. \tag{4.4}$$

- *Case III.* State transition from $(l_1, l_2) = (0, 1)$ to $(0, 0)$

  In this case, there is no arrival of either HP or LP user (with probability $\mathcal{D}_0$). The server completes service for one LP user and starts vacation (with probability $\mathbf{s}_2\delta$) since the buffer becomes empty after the service. In this case, the submatrix for transitions can be represented as

$$\mathcal{A}_0^{10} = \begin{bmatrix} 0 & \mathcal{D}_0 \otimes (\mathbf{s}_2\delta) \\ 0 & 0 \end{bmatrix}. \tag{4.5}$$

- *Case IV.* State transition from $(l_1, l_2) = (0, l_2')$ to $(0, l_2' + 1)$, $l_2' \in [1, k-1]$

  This case considers the events that there is one new arrival of the LP user and no service is finished. The unfinished service may result from i) incomplete service if the server is in service mode, ii) incomplete vacation, or iii) a vacation ends and service starts for the LP users. We can summarize the transition submatrix in case VI as

$$\mathcal{A}_0^{0} = \begin{bmatrix} \mathcal{D}_{0,1} \otimes \mathcal{S}_2 & 0 \\ \mathcal{D}_{0,1} \otimes (\mathbf{v}\beta_2) & \mathcal{D}_{0,1} \otimes \mathcal{V} \end{bmatrix}. \tag{4.6}$$

- *Case V.* System state stays at $(l_1, l_2) = (0, l_2')$

  This case consists of a few sub-scenarios: i) no arrival with incomplete service (with probability $\mathcal{D}_0 \otimes \mathcal{S}_2$), ii) one arrival from the LP users with one complete service (with probability $\mathcal{D}_{0,1} \otimes (\mathbf{s}_2\beta_2)$), iii) no arrival with vacation completion

and service starts for the LP users (with probability $\mathcal{D}_0 \otimes (\mathbf{v}\beta_2)$), or *iv*) no arrival with incomplete vacation (with probability $\mathcal{D}_0 \otimes \mathcal{V}$). In summary, the transition submatrix for case V is

$$\mathcal{A}_0^1 = \begin{bmatrix} \mathcal{D}_0 \otimes \mathcal{S}_2 + \mathcal{D}_{0,1} \otimes (\mathbf{s}_2\beta_2) & 0 \\ \mathcal{D}_0 \otimes (\mathbf{v}\beta_2) & \mathcal{D}_0 \otimes \mathcal{V} \end{bmatrix}. \tag{4.7}$$

- *Case VI.* State transition from $(l_1, l_2) = (0, l_2')$ to $(0, l_2' - 1)$, $l_2' \in [2, l_2]$

  In this case, the transitions mean a service for one LP user has been completed. Since there are LP users remaining in the buffer, the server will stay in the service mode. Obviously, there is no arrival in this case. Thus, we have

$$\mathcal{A}_0^2 = \begin{bmatrix} \mathcal{D}_0 \otimes (\mathbf{s}_2\beta_2) & 0 \\ 0 & 0 \end{bmatrix}. \tag{4.8}$$

- *Case VII.* The system stays at the boundary state $(l_1, l_2) = (0, k)$

  Since the system reaches buffer limit, the system state can remain unchanged if there is no arrival or a new arrival is blocked. The server can be in the service for the LP users, starting to serve the LP users after vacations, or in the middle of a vacation because there is no HP user in the system. By jointly considering all these scenarios, the transition submatrix becomes

$$\mathcal{A}_0^{11} = \begin{bmatrix} \mathcal{D}_{0,1} \otimes (\mathbf{s}_2\beta_2) + \mathcal{D} \otimes \mathcal{S}_2 & 0 \\ \mathcal{D} \otimes (\mathbf{v}\beta_2) & \mathcal{D} \otimes \mathcal{V} \end{bmatrix} \tag{4.9}$$

  where $\mathbf{s}_2 = \boldsymbol{e} - \mathcal{S}_2\boldsymbol{e}$ and $\mathbf{v} = \boldsymbol{e} - \mathcal{V}\boldsymbol{e}$.

In summary, the transition matrix $\mathcal{A}_0$ can be written as

$$\mathcal{A}_0 = \begin{bmatrix} \mathcal{A}_0^{00} & \mathcal{A}_0^{01} & & & \\ \mathcal{A}_0^{10} & \mathcal{A}_0^{1} & \mathcal{A}_0^{0} & & \\ & \mathcal{A}_0^{2} & \mathcal{A}_0^{1} & \mathcal{A}_0^{0} & \\ & & \ddots & \ddots & \ddots \\ & & & \mathcal{A}_0^{2} & \mathcal{A}_0^{11} \end{bmatrix} \tag{4.10}$$

where $\mathcal{A}_0$ is of $\mathbb{D}^{(k+1)\times(k+1)}$, and $\mathbb{D}$ denotes the dimension with respect to the number of submatrices.

Note that vacation termination may be triggered when the HP user arrives. Therefore, to explain the effect of this termination policy, we consider the state transition $\mathcal{B}_0^{01}$ in submatrix $\mathcal{B}_0$ as an example. $\mathcal{B}_0^{01}$ represents the state transition from $(l_1, l_2) = (0, 0)$ to $(1, 1)$. Since the event of an HP user arrival may happen just at the end of a vacation or during a vacation period, this results in the service of the HP user starting immediately or after vacation interruption. Such two scenarios can be jointly described as $(\mathbf{v} \otimes \beta_1 + \mathcal{V}\mathcal{Q} \otimes \beta_1)$. Thus, $\mathcal{B}_0^{01}$ can be derived as

$$\mathcal{B}_0^{01} = \begin{bmatrix} \mathcal{D}_{1,1} \otimes (\mathbf{v} + \mathcal{V}\mathcal{Q}) \otimes \beta_1 & 0 \end{bmatrix}. \tag{4.11}$$

The detailed formulae of all submatrices in transition matrix $\mathcal{P}$ are shown as follows.

$$\mathcal{B}_0 = \begin{bmatrix} \mathcal{B}_0^{00} & \mathcal{B}_0^{01} & & & \\ \mathcal{B}_0^{2} & \mathcal{B}_0^{1} & \mathcal{B}_0^{0} & & \\ & \ddots & \ddots & \ddots & \\ & & \mathcal{B}_0^{2} & \mathcal{B}_0^{1} & \\ & & & \mathcal{B}_0^{11} & \end{bmatrix} \tag{4.12}$$

where $\mathcal{B}_0 \in \mathbb{D}^{(k+1)\times k}$.

$$\mathcal{B}_0^{00} = \begin{bmatrix} \mathcal{D}_{1,0} \otimes (\mathbf{v} + \mathcal{V}\mathcal{Q}) \otimes \beta_1 & 0 \end{bmatrix}, \mathcal{B}_0^{01} = \begin{bmatrix} \mathcal{D}_{1,1} \otimes (\mathbf{v} + \mathcal{V}\mathcal{Q}) \otimes \beta_1 & 0 \end{bmatrix},$$

$$\mathcal{B}_0^{2} = \begin{bmatrix} \mathcal{D}_{1,0} \otimes (\mathbf{s}_2\beta_1) & 0 \\ 0 & \end{bmatrix}, \mathcal{B}_0^{1} = \begin{bmatrix} \mathcal{D}_{1,0} \otimes \mathcal{S}_2 + \mathcal{D}_{1,1} \otimes (\mathbf{s}_2\beta_1) & 0 \\ \mathcal{D}_{1,0} \otimes (\mathbf{v} + \mathcal{V}\mathcal{Q}) \otimes \beta_1 & \end{bmatrix},$$

$$\mathcal{B}_0^{0} = \begin{bmatrix} \mathcal{D}_{1,1} \otimes \mathcal{S}_2 & 0 \\ \mathcal{D}_{1,1} \otimes (\mathbf{v} + \mathcal{V}\mathcal{Q}) \otimes \beta_1 & \end{bmatrix}, \mathcal{B}_0^{11} = \begin{bmatrix} \mathcal{D}'_{10} \otimes (\mathbf{s}_2\beta_1) & 0 \\ 0 & \end{bmatrix},$$

and $\mathcal{D}'_{10} = \mathcal{D}_{1,0} + \mathcal{D}_{1,1}$. The derivation of $\mathcal{B}_0$ follows the same discipline as $\mathcal{A}_0$.

$$\mathcal{C}_0 = \begin{bmatrix} \mathcal{C}_0^{10} & \mathcal{C}_0^{00} & & \\ & \mathcal{C}_0^{1} & \mathcal{C}_0^{0} & \\ & & \ddots & \ddots \\ & & & \mathcal{C}_0^{1} & \mathcal{C}_0^{0} \end{bmatrix} \tag{4.13}$$

where $\mathcal{C}_0 \in \mathbb{D}^{k\times(k+1)}$, $\mathcal{C}_0^{10} = \begin{bmatrix} 0 & \mathcal{D}_0 \otimes (\mathbf{s}_1\delta) \end{bmatrix}$, $\mathcal{C}_0^{00} = \begin{bmatrix} \mathcal{D}_0 \otimes (\mathbf{s}_1\beta_2) & 0 \end{bmatrix}$,

$\mathcal{C}_0^{1} = \begin{bmatrix} \mathcal{D}_0 \otimes (\mathbf{s}_1\beta_2) & 0 \\ 0 & 0 \end{bmatrix}$, $C_0^{0} = \begin{bmatrix} D_{0,1} \otimes (\mathbf{s}_1\beta_2) & 0 \\ 0 & 0 \end{bmatrix}$, and $\mathbf{s}_1 = \boldsymbol{e} - S_1\boldsymbol{e}$.

$$\mathcal{A}_i = \begin{bmatrix} \mathcal{A}^{10} & \mathcal{A}^{00} & & & \\ \mathcal{A}^{2} & \mathcal{A}^{1} & \mathcal{A}^{0} & & \\ & \ddots & \ddots & \ddots & \\ & & \mathcal{A}^{2} & \mathcal{A}^{1} & \mathcal{A}^{0} \\ & & & \mathcal{A}^{2} & \mathcal{A}^{11} \end{bmatrix} \tag{4.14}$$

where $\mathcal{A}_i \in \mathbb{D}^{(k-i+1)\times(k-i+1)}$, $i = 1, 2, \ldots, k-1$,

$$\mathcal{A}^{10} = \begin{bmatrix} \mathcal{D}_0 \otimes \mathcal{S}_1 + \mathcal{D}_{1,0} \otimes (\mathbf{s}_1\beta_1), & 0 \end{bmatrix}, \mathcal{A}^{00} = \begin{bmatrix} \mathcal{D}_{0,1} \otimes \mathcal{S}_1 + \mathcal{D}_{1,1} \otimes (\mathbf{s}_1\beta_1), & 0 \end{bmatrix},$$

$$\mathcal{A}^{2} = \begin{bmatrix} \mathcal{D}_0 \otimes (\mathbf{s}_2\beta_1) & 0 \\ 0 & \end{bmatrix}, \mathcal{A}^{1} = \begin{bmatrix} \mathcal{D}_0 \otimes \mathcal{S}_1 + \mathcal{D}_{1,0} \otimes (\mathbf{s}_1\beta_1) & 0 \\ \mathcal{D}_0 \otimes \mathcal{S}_2 + \mathcal{D}_{0,1} \otimes (\mathbf{s}_2\beta_1) & \end{bmatrix},$$

$$\mathcal{A}^{0} = \begin{bmatrix} \mathcal{D}_{0,1} \otimes \mathcal{S}_1 + \mathcal{D}_{1,1} \otimes (\mathbf{s}_1\beta_1) & 0 \\ \mathcal{D}_{0,1} \otimes S_2 & \end{bmatrix}, \mathcal{A}^{11} = \begin{bmatrix} \mathcal{D}'_{10} \otimes S_1 + \mathcal{D}'_{10} \otimes (\mathbf{s}_1\beta_1) & 0 \\ \mathcal{D}'_{01} \otimes S_2 + \mathcal{D}_{0,1} \otimes (\mathbf{s}_2\beta_1) & \end{bmatrix},$$

where $\mathcal{D}'_{01} = \mathcal{D}_{0,0} + \mathcal{D}_{0,1}$. And $\mathcal{A}_k = \begin{bmatrix} \mathcal{D} \otimes \mathcal{S}_1 + \mathcal{D}'_{10} \otimes (\mathbf{s}_1\beta_1) & 0 \end{bmatrix}$.

$$\mathcal{B}_i = \begin{bmatrix} \mathcal{B}^{10} & \mathcal{B}^{00} & & & \\ \mathcal{B}^{2} & \mathcal{B}^{1} & \mathcal{B}^{0} & & \\ & \ddots & \ddots & \ddots & \\ & & & \mathcal{B}^{2} & \mathcal{B}^{11} \\ & & & & \mathcal{B}^{22} \end{bmatrix} \tag{4.15}$$

where $\mathcal{B}_i \in \mathbb{D}^{(k-i+1)\times(k-i)}$, $i = 1, 2, \ldots, k-2$,

$$\mathcal{B}^{10} = \begin{bmatrix} \mathcal{D}_{1,0} \otimes S_1 & 0 \end{bmatrix}, \mathcal{B}^{00} = \begin{bmatrix} \mathcal{D}_{1,1} \otimes S_1 & 0 \end{bmatrix},$$

$$\mathcal{B}^{2} = \begin{bmatrix} 0 & \\ \mathcal{D}_{1,0} \otimes (\mathbf{s}_2\beta_1) & 0 \end{bmatrix}, \mathcal{B}^{1} = \begin{bmatrix} \mathcal{D}_{1,0} \otimes \mathcal{S}_1 & 0 \\ \mathcal{D}_{1,0} \otimes \mathcal{S}_2 + \mathcal{D}_{1,1} \otimes (\mathbf{s}_2\beta_1) & \end{bmatrix},$$

$$\mathcal{B}^{0} = \begin{bmatrix} \mathcal{D}_{1,1} \otimes \mathcal{S}_1 & 0 \\ \mathcal{D}_{1,1} \otimes S_2 & \end{bmatrix}, \mathcal{B}^{11} = \begin{bmatrix} \mathcal{D}'_{10} \otimes \mathcal{S}_1 & 0 \\ \mathcal{D}'_{10} \otimes \mathcal{S}_2 + \mathcal{D}'_{10} \otimes (\mathbf{s}_2\beta_1) & \end{bmatrix},$$

$$\mathcal{B}^{22} = \begin{bmatrix} 0 & \\ \mathcal{D}'_{10} \otimes (\mathbf{s}_2\beta_1) & 0 \end{bmatrix}, \text{ and } \mathcal{B}_{k-1} = \begin{bmatrix} \mathcal{D}'_{10} \otimes S_1 & 0 \\ \mathcal{D}'_{10} \otimes (\mathbf{s}_2\beta_1) & \\ & 0 \end{bmatrix}.$$

$$\mathcal{C}_i = \begin{bmatrix} \mathcal{C}^{10} & \mathcal{C}^{00} & & \\ & \mathcal{C}^{1} & \mathcal{C}^{0} & \\ & & \ddots & \ddots \\ & & & \mathcal{C}^{1} & \mathcal{C}^{0} \end{bmatrix} \tag{4.16}$$

where $\mathcal{C}_i \in \mathbb{D}^{(k-i)\times(k-i+1)}$, $i = 1, 2, \ldots, k-2$,

$$\mathcal{C}^{10} = \begin{bmatrix} \mathcal{D}_0 \otimes (\mathbf{s}_1\beta_1) & 0 \end{bmatrix}, \mathcal{C}^{00} = \begin{bmatrix} \mathcal{D}_{0,1} \otimes (\mathbf{s}_1\beta_1) & 0 \end{bmatrix},$$

$$\mathcal{C}^{1} = \begin{bmatrix} \mathcal{D}_{0,0} \otimes (\mathbf{s}_1\beta_1) & 0 \\ & \\ 0 & \end{bmatrix}, \mathcal{C}^{0} = \begin{bmatrix} \mathcal{D}_{0,1} \otimes (\mathbf{s}_1\beta_1) & \\ & \\ 0 & 0 \end{bmatrix}.$$

And $\mathcal{C}_{k-1} = \begin{bmatrix} \mathcal{C}^{10} & \mathcal{C}^{00} \end{bmatrix}$.

4.2.2 Steady State Distribution

The system is stable only if the traffic intensity $\rho = \frac{\lambda_1}{\mu_1} + \frac{\lambda_2}{\mu_2} < 1$.

Let $x_{i,j}$ denote the probability that there are i HP and j LP users in the system. Note that $i + j \leqslant k$. We further define the finite stationary probability vector $x =$

$\{x_0, x_1, x_2, ..., x_k\}$ as

$$x = \begin{cases} x_0 = \{x_{0,0}, x_{0,1}, x_{0,2}, ..., x_{0,k}\} \\ x_i = \{x_{i,0}, x_{i,1}, x_{i,2}, ..., x_{i,(k-1)}\}, \quad 1 \leqslant i \leqslant k-1. \\ x_k = x_{k,0} \end{cases}$$

Then, the stationary probability vector x satisfies the following equation set

$$\begin{cases} x\left(\mathcal{P} - \mathcal{I}\right) = 0 \\ x\boldsymbol{e} = 1 \end{cases} \tag{4.17}$$

where $\mathcal{I}$ is an identity matrix of appropriate finite dimension. We adopt the method of Gaussian elimination with block state reduction [64] to solve x. We first reformat transition probability matrix $\mathcal{P}$ by removing n rows from the bottom as follows

$$\mathcal{P}_{k-n} = \begin{bmatrix} \mathcal{A}_0 & \mathcal{B}_0 & & & \\ \mathcal{C}_1 & \mathcal{A}_1 & \mathcal{B}_1 & & \\ & \mathcal{C}_2 & \mathcal{A}_2 & \mathcal{B}_2 & \\ & & \ddots & \ddots & \ddots \\ & & & \mathcal{C}_{k-n} & \mathcal{F}_{k-n} \end{bmatrix}, 0 \leqslant n \leqslant k-1 \tag{4.18}$$

where $\mathcal{F}_k = \mathcal{A}_k$ and

$$\mathcal{F}_{k-n} = \mathcal{A}_{k-n} + \mathcal{B}_{k-n}\left(\mathcal{I} - \mathcal{A}_{k-n+1}\right)^{-1}\mathcal{C}_{k-n+1},\ 1 \leqslant n \leqslant k. \tag{4.19}$$

By partitioning $\mathcal{P}_{k-n}$ into two blocks $\{0\}$ and $\{1, 2, ..., k-n\}$, the matrix $\mathcal{P}$ can be decomposed into four submatrices as

$$\mathcal{P}_{k-n} = \begin{bmatrix} \mathcal{P}_{00} & \mathcal{P}_{0\Lambda} \\ \mathcal{P}_{\Lambda 0} & \mathcal{P}_{\Lambda\Lambda} \end{bmatrix}, \tag{4.20}$$

where $\mathcal{P}_{\Lambda\Lambda}$ is the submatrix of $\mathcal{P}$ in the state space $\{1, 2, ..., k-n\}$. We similarly represent x as (x_0, x_Λ) with $x_\Lambda = [x_1, x_2..., x_{k-n}]$. From (4.17), we have

$$\begin{cases} x_0 = x_\Lambda \mathcal{P}_{\Lambda 0}(\mathcal{I} - \mathcal{P}_{00})^{-1} \\ x_\Lambda = x_\Lambda \mathcal{P}^1_{k-n} \end{cases} \tag{4.21}$$

where x_0 can be derived in a function of $\{x_1, ..., x_{k-n}\}$. $\mathcal{P}^1_{k-n}$ is the transition matrix of the censored Markov chain on the state space $\{1, ..., k-n\}$, and can be determined by $\mathcal{P}^1_{k-n} = \mathcal{P}_{\Lambda\Lambda} + \mathcal{P}_{\Lambda 0}(\mathcal{I} - \mathcal{P}_{00})^{-1}\mathcal{P}_{0\Lambda}$. Also, the stationary vector of $\mathcal{P}^1_{k-n}$ is proportional to x_Λ, i.e., $x_\Lambda \mathcal{P}^1_{k-n} = x_\Lambda$. Following the similar procedure, we can further partition $\mathcal{P}_{\Lambda\Lambda}$ in state space $\{1, ..., k-n\}$ into subsets $\{1\}$ and $\{2, ..., k-n\}$, and obtain an expression for x_1 in function of $\{x_2, ..., x_{k-n}\}$ with $\mathcal{P}^2_{k-n} = \mathcal{P}'_{\Lambda\Lambda} + \mathcal{P}_{\Lambda 1}(\mathcal{I} - \mathcal{P}_{11})^{-1}\mathcal{P}_{1\Lambda}$. In summary, from (4.19) and (4.21), the overall steady-state distribution can be calculated recursively as

$$\begin{cases} x_0 = x_0 \mathcal{F}_0 \\ x_{k-n} = x_{k-n-1}\mathcal{B}_{k-n-1}\left(\mathcal{I} - \mathcal{F}_{k-n}\right)^{-1}, \quad 0 \leqslant n \leqslant k-1 \end{cases} \tag{4.22}$$

with $\sum_{n=0}^{k} x_n = 1$.

4.2.3 Performance Measures

In this section, we evaluate system performance in terms of *queue length*, *user waiting time*, *blocking probability* and *system busy cycle*.

Queue length: Define y_n, $n = \{0, 1, ..., k\}$, as the marginal probability density of finding n users in the system. We have $y_n = \{x_{i,j} | i + j = n\}\boldsymbol{e}$, and $\boldsymbol{e}$ is a column vector of ones with an appropriate dimension. Then, the average number of users in

the system is

$$\overline{\mathcal{L}} = \sum_{n=0}^{k} n y_n. \tag{4.23}$$

The probability that there are i HP users in the system is

$$\mathcal{L}_1 = \sum_{j=0}^{k-i} x_{i,j}\boldsymbol{e}, 0 \leqslant i \leqslant k. \tag{4.24}$$

When considering the probability of j LP users in the system, we need to discuss two scenarios separately. *i*) If there is no LP user in the system, the situation becomes either there is no HP user in the system or there exist HP users with one HP user in service. *ii*) If the number of the LP users belongs to $\{1, 2, ..., k\}$, the situation further involves three scenarios: *a*) there is no HP user in the system and the server is on vacation, *b*) an HP user is in service, and *c*) an HP user arrives but the service to one LP user is incomplete. Thus, the probability that there are j LP users in the system can be derived as

$$\mathcal{L}_2 = \begin{cases} x_{0,0}\boldsymbol{e} + \sum_{i=1}^{k} x_{i,0}\boldsymbol{e}_1, & j = 0 \\ x_{0,j}\boldsymbol{e} + \sum_{i=1}^{k-j} x_{i,j}\boldsymbol{e}_1 + \sum_{i=1}^{k-j} x_{i,(j-1)}\boldsymbol{e}_2, & 1 \leqslant j \leqslant k \end{cases} \tag{4.25}$$

where $\boldsymbol{e}_1 = (\boldsymbol{e}\ \boldsymbol{e}_0)^{tr}$ and $\boldsymbol{e}_2 = (\boldsymbol{e}_0\, \boldsymbol{e})^{tr}$ indicate one HP or LP user in service, respectively. $\boldsymbol{e}_0$ is a column vector of all zeros. From (4.24) and (4.25), the average numbers of the HP and LP users in the system are $\bar{\mathcal{L}}_1 = \sum_{i=0}^{k} i\mathcal{L}_1$ and $\bar{\mathcal{L}}_2 = \sum_{j=0}^{k} j\mathcal{L}_2$, respectively.

User waiting time: Define $\overline{\mathcal{W}}$ as the users' average waiting time in the system. Then, we have $\overline{\mathcal{W}} = \overline{\mathcal{L}}/\lambda$ by applying Little's law. Similarly, the average waiting time of the HP and LP users can be calculated as $\overline{\mathcal{W}}_1 = \overline{\mathcal{L}}_1/\lambda_1$ and $\overline{\mathcal{W}}_2 = \overline{\mathcal{L}}_2/\lambda_2$, respectively.

Blocking probability: A newly arrived HP user can be blocked if *i*) the system buffer is full with k users (no matter what types of users) and the current service is in phase $\mathcal{S}_1$ or $\mathcal{S}_2$, or *ii*) the system buffer is full with k LP users and the server is on vacation. Let $\mathcal{D}_{21} = \mathcal{D}_{1,0} + \mathcal{D}_{1,1}$ and $\mathcal{S}_{12} = \mathcal{S}_1 + \mathcal{S}_2$. Thus, the blocking probability of the HP user can be calculated as

$$\mathcal{P}_{\mathcal{B}_1} = \frac{1}{\lambda_1} \begin{cases} x_{i,j} \left[\mathcal{D}_{21} \otimes \mathcal{S}_{12}\right], & i + j = k \\ x_{i,j} \left[\mathcal{D}_{21} \otimes \mathcal{V}\right], & i = 0, j = k \end{cases} \tag{4.26}$$

When an LP user arrives, the blocking event happens in three cases. *i*) If only one position in the buffer is available, i.e., the queue length is $i + j = k - 1$, and there is an HP user arriving at the same time, the HP user has a priority to occupy the only position, *ii*) The buffer is full, $i + j = k$, and the system is in service, and *iii*) There are k LP users in the system, $i = 0, j = k$, and the server is on vacation. Thus, the blocking probability of the LP user can be calculated as

$$\mathcal{P}_{B_2} = \frac{1}{\lambda_2} \begin{cases} x_{i,j} \left[\mathcal{D}_{1,1} \otimes \mathcal{S}_{12}\right] & i + j = k - 1 \\ x_{i,j} \left[\mathcal{D}_{12} \otimes \mathcal{S}_{12} + \mathcal{D}_{1,1} \otimes (\mathbf{s}\beta)_{12}\right] & i + j = k \\ x_{i,j} \left[\mathcal{D}_{12} \otimes \mathcal{V}\right] & i = 0, j = k \end{cases} \tag{4.27}$$

where $S_{12} = S_1 + S_2$, $\mathcal{D}_{12} = \mathcal{D}_{0,1} + \mathcal{D}_{1,1}$, and $(\mathbf{s}\beta)_{12} = \mathbf{s}_1\beta_1 + \mathbf{s}_2\beta_2$.

System busy cycle: Since the system consists of alternative vacation and service periods, it is essential to analyze such cyclic behavior for the purpose of system performance optimization. Let t_v and t_s denote the lengths of vacation and service periods, respectively. The system can leave the vacation mode when *i*) at least one LP user waiting in the buffer at the end of the current vacation, or *ii*) an HP user

arrival interrupts the vacation. The service period, t_s, starts immediately after the system leaves the vacation mode, and ends when the buffer becomes empty.

Definition 4 (*system busy cycle*)**.** A system busy cycle starts from the beginning of a vacation period when the buffer is empty, and ends when the next service period completes. Let t_b denote the length of the system busy cycle. It can be formulated as

$$t_b = t_v + t_s. \tag{4.28}$$

To derive the mean length of system busy cycle, we first calculate the mean vacation length. Let $v(t)$ denote the probability that the length of overall vacation period experiences t slots, and there are user arrivals in the last t' slots. Note that $v(t)$ is a joint probability of two scenarios: i) there is no user arrival and the buffer is empty in the first $t - t'$ slots, and ii) there are LP user arrivals during last t' slots. Note that the arrival of an HP user is not considered since the vacation will be interrupted with the HP user arrivals. Thus, based on the distribution of queue length, $v(t)$ can be formulated as

$$v(t) = \sum_{t'=0}^{t} \left(x_{0,0}\boldsymbol{e}\right)^{t-t'} \left(\sum_{j=1}^{k} x_{0,j}\boldsymbol{e}\right)^{t'}. \tag{4.29}$$

For the mean service period, let $\mathcal{G}(l,t)$ denote the probability in phase type that l users complete service within t slots. To calculate $\mathcal{G}(l,t)$, three cases have to be discussed.

- Case I: $l = 1, t \geqslant 1$. If one user of any type completes service in more than one slot, it means such service is not completed in the first $t-1$ slots and it is finished in the last slot. Thus, we have

$$\mathcal{G}^1(l,t) = \mathcal{S}_1^{t-1}\left(\mathbf{s}_1\beta_1\right) + \mathcal{S}_2^{t-1}\left(\mathbf{s}_2\beta_2\right). \tag{4.30}$$

- Case II: $l = t, t \geqslant 1$. If the server uses t slots to serve t users, apparently, each type of user completes its service in one slot. Thus, by considering l_1 HP users and l_2 LP users in the system, we have

$$\mathcal{G}^2(l,t) = (\mathbf{s}_1\beta_1)^{l_1} + (\mathbf{s}_2\beta_2)^{l_2}\,. \tag{4.31}$$

- Case III: $l \geqslant 2, t \geq l$. If the server uses t, $t > l$, slots to serve l users, in this case, any slot can be considered in the situation of either service completion or incomplete service. For example, in slot $t-1$, the server has to be in two situations: either the service of user $l-1$ is completed or service of user l is incomplete since only one slot left. Thus, $\mathcal{G}(l,t)$ in this case can be retrieved recursively, and at the end, it turns to either Case I or Case II. Thus, for Case III, we have

$$\mathcal{G}^3(l,t) = (\mathbf{s}\beta)_{12}\,\mathcal{G}(l-1,t-1) + \mathcal{S}_{12}\mathcal{G}(l,t-1). \tag{4.32}$$

In summary, by jointly considering the preceding three cases, the probability that l users complete service in t slots can be calculated as

$$\mathcal{G}(l,t) = \begin{cases} \mathcal{S}_1^{t-1}\,(\mathbf{s}_1\beta_1) + \mathcal{S}_2^{t-1}\,(\mathbf{s}_2\beta_2) & l = 1, t \geqslant 1 \\ (\mathbf{s}_1\beta_1)^{l_1} + (\mathbf{s}_2\beta_2)^{l_2} & l = t, t \geqslant 1 \\ (\mathbf{s}\beta)_{12}\,\mathcal{G}(l-1,t-1) + \mathcal{S}_{12}\mathcal{G}(l,t-1) & l \geqslant 2, t > l \end{cases} \tag{4.33}$$

Let $\mathcal{H}(l,t)$ denote the probability in phase type that a service period lasts t slots given there are l users at the beginning of the service period. The derivation of $\mathcal{H}(l,t)$ needs to consider the following two cases.

- Case I: There is no user arrival during the service period.

  This case is simple because only the initial users at the beginning of service need to be considered. Such procedure can be described as $\mathcal{H}(l,t) = \mathcal{G}(l,t) \otimes \mathcal{D}_0^t$, where $\mathcal{D}_0^t$ means there is no arrival in t slots.

- Case II: There exist new arrivals within the service period.

  In this case, assume that at time t', the server finishes the service to the initiated l users and there are l' new arrivals in t'. Such procedure can be described as $\mathcal{G}(l,t') \otimes \sum_{l'=1}^{2t'} d(l',t')$, where $d(l',t')$ indicates the probability that there are l' user arrivals during t' slots, and can be calculated by

$$d(l',t') = \sum_{i=1}^{l'} \binom{t'}{i} \binom{t'-i}{\frac{l'-i}{2}} \left(\mathcal{D}_{0,1} + \mathcal{D}_{1,0}\right)^i \mathcal{D}_0^{t-\frac{l'+i}{2}} \mathcal{D}_{1,1}^{\frac{l'-i}{2}}. \tag{4.34}$$

  Note that the maximum number of user arrivals is $2t'$, since in one slot, there are at most two user arrivals with one for each type. From t', we can assume a beginning of a new service period, which has l' users at the initial state. We can repeat the similar procedure till the end of the service period. Therefore, $\mathcal{H}^l(t)$ can be formulated as

$$\mathcal{H}(l,t) = \mathcal{G}(l,t) \otimes \mathcal{D}_0^t + \sum_{t'=1}^{t-1} \left(\mathcal{G}(l,t') \otimes \sum_{l'=1}^{2t'} d(l',t') \right) \mathcal{H}(l', t-t'). \tag{4.35}$$

Let $h(l,t)$ denote the probability that a service period initiated by l users lasts t slots. Then, we have

$$h(l,t) = \begin{cases} \psi \mathcal{H}(l,t)\boldsymbol{e}, & l \leqslant \min(t,k) \\ 0 & l > t \end{cases} \tag{4.36}$$

where ψ is the stationary probability vector that satisfying $\psi(\mathcal{I} - \mathcal{H}(l,t)) = \psi$, and $\psi \boldsymbol{e} = 1$.

From (4.29) and (4.36), the mean values of service and vacation periods can be calculated as $\overline{t_s} = \sum_{t=l}^{\infty} th(l,t)$, and $\overline{t_v} = \sum_{t=0}^{\infty} tv(t)$, respectively. Then, from (4.28), the mean system busy cycle equals $\overline{t_b} = \overline{t_v} + \overline{t_s}$.

4.3 Strategy of Adaptive Service Rate and Vacation Length

In this section, we first explain the tradeoff between system parameters and performance. Then, we propose the ASV method to maximize the HeNB energy efficiency and further introduce the one-step look-ahead method to reduce the computational complexity.

4.3.1 Relationship between System Parameters and Performance

Based on the previous queueing analysis, the HeNB operates in consecutive busy cycles with alternative service and vacation periods. Therefore, the service rate and the vacation length are two key factors determining system performance in terms of energy efficiency and QoS. We summarize the relationship among *vacation length*, *service rate*, *system energy efficiency* and *QoS requirements* as follows

- *Vacation length* v.s. *Energy efficiency*: In order to achieve enhanced energy efficiency, in any busy cycle, it is essential to have a long and uninterrupted

vacation instead of multiple short vacations, since a long vacation can save energy consumption in the listening periods.

- *Vacation length* v.s. *QoS*: A long vacation may result in users' waiting time over maximum delay requirement or exceeding the blocking probability threshold. A long vacation also has a high chance to be interrupted by the arrival of the HP users. In addition, changing vacation length may affect the initial state, i.e., the number of users in the buffer, of the next service period so as to cause a heavier burden to the following service period.

- *Service rate* v.s. *Vacation* and *Energy efficiency*: When the HeNB enters a service period, it may adopt a high service rate to quickly clear the buffer so that the HeNB enters a long vacation. However, choosing high service rate consumes more energy, which may degrade the energy efficiency.

- *Service rate* v.s *Vacation* and *QoS*: In the condition that the HeNB is in the service period and the HP user is expected to arrive soon, it is wise to adopt a low service rate so that the service of the incoming HP user can be completed before the initiation of a vacation to avoid a quick vacation interruption. However, reducing service rate may increase users' waiting time and blocking rate.

4.3.2 Maximizing System Energy Efficiency

To balance the tradeoff on system performance, in this section, our objective is to maximize the system energy efficiency and satisfy QoS requirements at the same time by adjusting the service rate and vacation length. Since the system behavior

is running with alternative vacation and service periods, we can define the system energy efficiency, denoted by E_{sys} as

$$E_{\text{sys}} = \sum_{n=1}^{n_b} \mathcal{E}_n^b(r_n, v_n) \tag{4.37}$$

where n_b denotes the number of system busy cycles. $\mathcal{E}_n^b(r_n, v_n)$ denotes system energy efficiency for each busy cycle n, which is the function of selected service rate r_n and vacation length v_n. In any busy cycle, $\mathcal{E}^b(r, v)$ can be formulated as a logarithmic function of system throughput over the total energy consumption on both service and vacation periods. Specifically,

$$\mathcal{E}^b(r, v) = \log\left(\frac{\sum_{i=1}^{l_1+l_2} r_i t_i}{\sum_{i=1}^{l_1+l_2} p_i + \sum_{\alpha=1}^{N_v^b} (v_\alpha \cdot p_v + p_l)}\right) \tag{4.38}$$

where r_i, p_i, t_i denote the service rate, the corresponding energy consumption and the number of slots in service for user i, respectively. N_v^b denotes the total number of consecutive vacations in a busy cycle. v_α denotes the vacation length in the number of slots. Note that although there is no system throughput when the server is on vacation, the system still consumes some power in vacation and listening periods, which are denoted as p_v and p_l, respectively. Commonly, $p_i \gg p_l \gg p_v \geqslant 0$. Here, we adopt a logarithmic function for energy efficiency to avoid potential marginal solutions.

Then, the problem of maximizing system energy efficiency can be formulated as

$$\max_{\{r\},\{v\}} \quad E_{\text{sys}} = \sum_{n=1}^{n_b} \mathcal{E}_n^b(r, v) \tag{4.39}$$

$$\text{s.t.} \quad \mathcal{P}_{\mathcal{B}_1} \leqslant \eta_1, \mathcal{P}_{\mathcal{B}_2} \leqslant \eta_2 \tag{4.40}$$

$$\overline{\mathcal{W}}_1 \leqslant \omega_1, \overline{\mathcal{W}}_2 \leqslant \omega_2 \tag{4.41}$$

$$\sum_{i=1}^{l_1+l_2} t_i \geqslant l_1 + l_2, l_1 + l_2 = 0, 1, \ldots, k \tag{4.42}$$

$$\{r\} \in \mathbb{R}, \{v\} \in \mathbb{V} \tag{4.43}$$

where constraint (4.40) guarantees that blocking probability of the HP (LP) user is not greater than a predefined threshold η_1 (η_2). Constraint (4.41) guarantees that the waiting time of the HP (LP) users is less than a predefined maximum delay ω_1 (ω_2). Constraint (4.42) limits that the number of slots in service must be greater than the number of users served since at least one slot is needed to serve a user.

4.3.3 The ASV Method with Dual Decomposition Solution

We try to solve the problem of maximization defined in (4.39) $\sim$ (4.43) by achieving the maximum HeNB energy efficiency in each busy cycle. Thus, we rewrite the problem (4.39) in each busy cycle as

$$\mathbf{P}^*: \max_{\{r\},\{v\}} \log\left(\sum_{i=1}^{l_1+l_2} r_i t_i\right) - \log\left(\sum_{i=1}^{l_1+l_2} p_i + \sum_{\alpha=1}^{N_v^b}(v_\alpha p_v + p_l)\right) \tag{4.44}$$

$$\text{s.t.} \quad (4.40) \sim (4.43).$$

Next, we adopt the Lagrangian dual decomposition method to relax the coupled constraints by introducing a Lagrange multiplier ξ. The dual problem is:

$$\mathbf{D}^*: \min_{\xi} \quad f_r(\xi) + g_v(\xi) \tag{4.45}$$

where

$$f_r(\xi) = \begin{cases} \max\limits_{\{r\}} & \log\left(\sum_{i=1}^{l_1+l_2} r_i t_i\right) - \xi \\ \text{s.t.} & \mathcal{P}_{\mathcal{B}_1} \leqslant \eta_1, \overline{\mathcal{W}}_1 \leqslant \omega_1 \\ & \sum_{i=1}^{l_1+l_2} t_i \geqslant l_1 + l_2, l_1 + l_2 = 0, 1, \ldots, k \\ & \{r\} \in \mathbb{R} \end{cases} \tag{4.46}$$

$$g_v(\xi) = \begin{cases} \max\limits_{\{v\}} & \xi - \log\left(\sum_{i=1}^{l_1+l_2} p_i + \sum_{\alpha=1}^{N_v^b} (v_\alpha p_v + p_l)\right) \\ \text{s.t.} & \mathcal{P}_{\mathcal{B}_2} \leqslant \eta_2, \overline{\mathcal{W}}_2 \leqslant \omega_2 \\ & l_1 + l_2 \leqslant k \\ & \{v\} \in \mathbb{V} \end{cases} \tag{4.47}$$

When the constraints in the problem (4.44) are all linear equalities and inequalities, based on *Slater's condition* (or *constraint qualification*) [65], the problem (4.44) holds *strong duality*. Thus, the optimal value of problem (4.44), denoted by P^*, and the optimal value of Lagrange dual problem (4.45), denoted by D^*, must have same optimal value, which means the optimal duality gap is zero, i.e., $\mathrm{D}^* = \mathrm{P}^*$. Since the constraints (4.40)$\sim$(4.41) are non-linear inequalities, the problem (4.44) holds *weak duality* based on *Slater's condition* (or *constraint qualification*) [65]. Thus, the optimal value of problem (4.44), denoted by P^*, and the optimal value of Lagrange dual problem (4.45), denoted by D^*, hold $\mathrm{D}^* \leqslant \mathrm{P}^*$. Since the dual problem (4.45) is always convex, we can always find the best lower bound on the primal problem (4.44). Therefore, given the dual optimal ξ^*, we can at least obtain the near optimal solution by solving the decoupled optimization problems (4.46) and (4.47) separately without coordination among the service and vacation periods.

To solve the dual maximization problems, we propose a solution consisting of inner and outer optimization, which captures the special structure in system behavior, i.e., vacation periods followed by a service period in each busy cycle. The adaptive service rate from (4.46) is derived in the inner optimization, while the adaptive vacation length from (4.47) can be obtained in the outer optimization.

The basic idea of the proposed ASV method can be described by the following three steps.

- *Step 1.* In inner optimization, by considering all possible initial states, the local optimal sets of service rates r^*, denoted by $\bar{R}$, can be derived by adopting dynamic programming.

- *Step 2.* In outer optimization, when a vacation length is selected, the queue length at the time when the vacation ends and the service starts can be estimated based on the queueing analysis in Section 4.2. Then, by applying $\bar{R}$ from *step 1*, the optimal vacation length v^*, denoted by $\bar{V}$, can be obtained and the corresponding r^* can be selected according to the related initial states. If there are multiple global optimums, the system will choose any one of them.

- *Step 3.* By applying *step 1* and *step 2* in each busy cycle, the system is expected to achieve maximum energy efficiency in a long run.

We discuss the detailed solution procedures in the following subsections.

Inner Optimization

Define $t'_n \in \mathbb{T}$, $n = 0, 1, ..., \mathcal{T}-1$, as the time epochs that a service starts for an HP (LP) user, where $\mathbb{T}$ represents the total HeNB operational time. At each time epoch

t'_n, the HeNB makes an action to select a service rate $r_i \in \mathbb{R}$. Note that the slots required in service t_i and the power consumption p_i are the functions of r_i. Define a utility function $u(r_i)$ of energy efficiency for serving user i as

$$u\left(r_i\right) = \frac{r_i t_i}{p_i}. \tag{4.48}$$

Then, at time epoch t'_n, the HeNB determines r_i^* as

$$r_i^* = \arg\max_{r_i} U(t'_n) \tag{4.49}$$

where

$$U(t'_n) := \max_{r_i \in \mathbb{R}, t'_n \in \mathbb{T}} u_{t'_n}\left(r_i\right). \tag{4.50}$$

Obviously, the optimization of r_i is a multistage decision making problem, which can be addressed through the following dynamic programming approach.

Define a decision sequence $\mathbb{A} = \{a_n\}$, $n = \{0, 1, \ldots, \mathcal{T} - 1\}$, representing actions of selecting appropriate service rate r_i. If this action space consists of a feasible solution to problem $U(t'_n)$, $\mathbb{A}$ will be deemed feasible at state $t'_n \in \mathbb{T}$. Similarly, if $\mathbb{A}$ constitutes an optimal solution to problem $U(t'_n)$, such decision sequence will be deemed optimal with respect to state t'_n.

Theorem 4.1. *At each time epoch $t_{n'} \in \mathbb{T}$, $U(t'_n)$ has at least one optimal solution.*

Proof. The proposition clearly holds for each t'_n, since all admissible service rate $\mathbb{R}$ is compact, i.e., finite and non-empty. Based on *Weierstrass* extreme value theorem [66], for each t'_n, $r_i \in \mathbb{R}$ is in a closed and bounded interval. Thus, the continuous function of $r_i \in \mathbb{R}$ must attain the maximum and minimum values, each at least once. Thus, $U(t'_n)$ has at least one optimal solution. □

Theorem 4.2. *Given policy $\pi(t'_n, a_n)$ for any pair of time epoch and action (t'_n, a_n), $t'_n \in \mathbb{T}$ and $a_n \in \mathbb{A}$, there exists a policy π^* that is optimal for $U(t'_n)$.*

Proof. Let $\mathbb{A}^* = \left(a_0^*, a_1^* \ldots, a_{\mathcal{T}-1}^*\right)$ be any optimal solutions to $U(t'_n)$. Define a policy $\pi(t'_n) = (a_n^*)$, $a_n^* \in \mathbb{A}^*, t'_n \in \mathbb{T}$. Note that applying this policy can always generate the sequence of decisions $\left(a_0^*, a_1^* \ldots, a_{\mathcal{T}-1}^*\right)$ at time epochs $0, 1, \ldots, \mathcal{T}-1$. Since $\left(a_0^*, a_1^* \ldots, a_{\mathcal{T}-1}^*\right)$ is an optimal solution to problem $U(t'_n)$, it follows that π^* is an optimal policy for the problem $U(t'_n)$. □

Since the action depends only on current state without relevance to the prior history, $\pi(t'_{n+1}, a_n)$ is only determined by $\pi(t'_n, a_n)$, which satisfies Markovian properties. For any policy π, define an evaluation function $\varphi_n^\pi(l'_n)$ for stage $n = \{0, 1, \ldots, \mathcal{T}\}$, which evaluates the system energy efficiency when there are l' users in the system. Particularly, at final stage $\mathcal{T}$, the service period is completed and there are no users in the system such that $\varphi_{\mathcal{T}}^\pi(l'_{\mathcal{T}}) = 0$. Thus, the evaluation function at any previous stage can be computed via backward induction by using Bellman equation as

$$\varphi_n^\pi(l'_n) = \begin{cases} \max\limits_{\pi(t'_n, a_n)} & \left(u_n\left(r_i\right) + \gamma^n \cdot \varphi_{n+1}^\pi(l'_{n+1})\right) \\ \text{s.t.} & l'_{n+1} = l'_n + d'_n - 1 > 0, n \in [0, \mathcal{T}-1] \\ & (4.40) \sim (4.41) \end{cases} \tag{4.51}$$

where d'_n denotes the number of user arrivals at stage n, and γ, $0 < \gamma \leqslant 1$, denotes the discount factor. $\gamma = 1$ represents the undiscounted case.

Let $\Pi^* = \{\pi^*(t'_n, a_n)\}$ denote optimal policies. Once maximum $\varphi_n^\pi(l_n)$ has been found, an optimal action a_n^* for stage n can be determined with the underlying optimal

policy $\pi^*(t_n, a_n^*)$ as

$$\pi^*(t'_n, a_n^*) = \arg \max_{\pi(t'_n, a_n)} \varphi_n^{\pi}(l'_n). \tag{4.52}$$

Proposition 1. *The computational complexity of the inner optimization is* $\mathcal{O}(k\mathcal{N}\mathcal{T})$.

Proof. The initial state of any service period consists of two scenarios: *i*) Due to vacation interruption, the system has one HP user and $l'-1$ LP users, and *ii*) When vacation ends without interruption, the system has l' LP users. By applying exhaustive search, maximum $2 \cdot k$ rounds are needed for traversing all conditions since $1 \leqslant l' \leqslant k$. The computation should also consider all the admissible service rates $\mathcal{N}$ and all states $\mathcal{T}$, i.e., the number of all served users, which results in the total computational complexity equal to $\mathcal{O}(k\mathcal{N}\mathcal{T})$. □

Outer Optimization

In outer optimization, to derive the optimal vacation length, we adopt an online reinforcement learning algorithm, called Q-learning algorithm [67]. Since there may be multiple vacations in each busy cycle, we define time epochs at the beginning of each vacation as t'_ϵ, $\epsilon = 1, 2, ..., N_v^b$ with $N_v^b < \infty$. Let $\zeta_\epsilon(v_i)$, $\epsilon = 1, 2, ..., N_v^b$, denote all admissible values of vacation length at each stage ε. Recall that $\{v_i\} \in \mathbb{V}$, $i = 1, 2, ..., \mathcal{M}$.

Assume that the system has an approximated Q-value denoted by $Q_\epsilon(t'_\epsilon, v_\epsilon)$. It means the system selects vacation length $v_\epsilon \in \mathbb{V}$ at time epoch t'_ϵ, which generates a value $Q_\epsilon(t'_\epsilon, v_\epsilon)$. Obviously, if an optimal Q-value is achievable, the optimal vacation length can also be determined. At initial stage $\epsilon = 1$, the vacation length can be selected by adopting random walk in the admissible set $\mathbb{V}$, i.e., $v_\epsilon \in \mathbb{V}$. Then, the

selected vacation length v_ϵ can trigger the system to generate cost function c, which is formulated as

$$c_{t'_\epsilon}(v_\epsilon) = \begin{cases} \mathcal{E}^b(r_\epsilon(l_\epsilon^{ini}), v_\epsilon), & l_\epsilon^{ini} \in [1, k] \\ 0, & l_\epsilon^{ini} = 0 \end{cases} \tag{4.53}$$

where $\mathcal{E}^b$ refers to the energy efficiency defined in (4.39). $r_\epsilon \in \bar{R}$ is the optimal set of service rates derived from the inner optimization with corresponding l_ϵ^{ini}, which denotes the initial number of users at the beginning of next service period. Particularly, the system starts a new vacation when $l_\epsilon^{ini} = 0$ with a zero cost.

Then, for multiple vacation periods in one busy cycle, the optimization problem can be formulated as

$$\max_{v_\epsilon \in \mathbb{V}} \sum_{\epsilon=1}^{N_v^b} c_{t'_\epsilon}(v_\epsilon) \tag{4.54}$$

$$\text{s.t. } \mathcal{P}_{\mathcal{B}_2} \leqslant \eta_2, \overline{\mathcal{W}}_2 \leqslant \omega_2.$$

According to Q-learning algorithm, let $\alpha_\epsilon\left(t'_\epsilon, v_\epsilon\right), 0 < \alpha_\epsilon \leqslant 1$, denote the learning rate, which can update Q-value at each iteration based on the system requirements. When the system state changes to $t'_{\epsilon+1}$, then $Q_{\epsilon+1}$ can be calculated as

$$\begin{aligned} Q_{\epsilon+1}\left(t'_\epsilon, v_\epsilon\right) &= \left(1 - \alpha_\epsilon + \beta_\epsilon\right) Q_\epsilon\left(t'_\epsilon, v_\epsilon\right) \\ &\quad + \left(\alpha_\epsilon - \beta_\epsilon\right)\left[c_{t'_\epsilon}(v_\epsilon) + \gamma \Theta_\epsilon(t'_{\epsilon+1}, v_{\epsilon+1})\right] \end{aligned} \tag{4.55}$$

where $\beta_\epsilon \in [0, 1)$ denotes the rate of penalty due to vacation interruption. $\beta_\epsilon = 0$ if vacation is not interrupted. Note that $0 \leqslant \beta_\epsilon < \alpha_\epsilon < 1$. $\Theta_\epsilon(t'_{\epsilon+1} + 1, v_{\epsilon+1})$ indicates the estimated optimal Q-value in the next state $t_{\epsilon+1}$ and can be calculated as

$$\Theta_\epsilon(t'_{\epsilon+1}, v_{\epsilon+1}) = \max_{\mathrm{v}_{\epsilon+1}} Q_\epsilon\left(t'_{\epsilon+1}, v_{\epsilon+1}\right). \tag{4.56}$$

$\gamma \in (0, 1]$ is a discount factor as described in the previous section. The learning rate α_ϵ determines the degree how current value obtained can override the old value. When $\alpha_\epsilon = 0$, the system will stop learning, while $\alpha_\epsilon \to 1$ enables the system to only accept the up-to-date value. Note that a constant learning rate is often adopted in practice.

Let $c^*_{\tau_\epsilon}$ and v^*_ϵ denote the desired optimal values. We have

$$c^*_{t'_\epsilon} = \max_{v_\epsilon} Q_\epsilon \left(t'_\epsilon, v_\epsilon\right), \tag{4.57}$$

$$v^*_\epsilon = \arg\max_{v_\epsilon} Q_\epsilon \left(t'_\epsilon, v_\epsilon\right). \tag{4.58}$$

Before the learning process starts, it returns an arbitrary fixed Q-value. This algorithm repeats at the beginning of each vacation period and ends when $t'_{\epsilon+1}$ is a final state. For all final states, $\Theta_\epsilon(t'_{\epsilon+1}, v_{\epsilon+1})$ stops updating and thus retains its initial value. Without loss of generality, we adopt $\Theta_\epsilon(t'_{\epsilon+1}, v_{\epsilon+1}) = 0$.

Proposition 2. *The computational complexity of the outer optimization is* $\mathcal{O}\left(\mathcal{M}N_v^b\right)$.

Proof. The outer optimization terminates after at most $\sum_{\epsilon=1}^{N_v^b} \sum_{i=1}^{\mathcal{M}} Q_\epsilon \left(t'_\epsilon, v_\epsilon\right)$ steps, which results in the computational complexity equal to $\mathcal{O}\left(\mathcal{M}N_v^b\right)$. □

4.3.4 One-step Look-ahead Method

In the outer optimization, since the total number of vacations in a busy cycle, N_v^b, is a small integer, the computational complexity mostly depends on the number of admissible vacation lengths $\mathcal{M}$, which can be determined by the mean vacation length based on the previous queueing analysis. However, in inner optimization, the underlying idea is to use backward recursion to approximate the optimality of r_i.

Therefore, if l' is large, which means in a service period, a lot of users are waiting for service, or if $l \to \infty$, backward recursion may involve considerable computational complexity. Thus, the inner optimization dominates the complexity of the overall solution procedure.

In order to reduce the computational complexity of the inner optimization, we propose an effective method, called look-ahead policy. In this policy, the system makes a decision, denoted by $\pi'(t'_n, a_n)$ at the beginning of service period. The one-step look-ahead policy is

$$\varphi_n^{\pi'}(l'_n) = \max_{\pi'} \left(u_n^{\pi'} + \Phi_{n+1}(l'_{n+1}, \pi') \right) \tag{4.59}$$

where $\Phi_{n+1}(l'_{n+1}, \pi') = \gamma^n \cdot \varphi_{n+1}^{\pi'}(l'_{n+1})$ is the approximated value function of the next stage. Define $u_n = \Phi_n$. Then, we have $u_{\mathcal{T}} = \Phi_{\mathcal{T}}$ at the final stage. Thus, by tracing back only one stage, the system can obtain a close to the maximum value based on (4.51) under identical system constraints.

When policy π' achieves the maximum value, Φ_{n+1} can be obtained on the basis of the one-step look-ahead approximation with respect to Φ_{n+2}. In other words, for all possible states t'_{n+1}, if using $\Phi_{n+1}(t'_{n+1})$ instead of $\Phi_{n+1}(l'_{n+1}, \pi')$, we obtain

$$\Phi_{n+1}(t'_{n+1}) = \max_{\pi'} \left(u_{n+1}^{\pi'} + \Phi_{n+2}(t'_{n+2}) \right) \tag{4.60}$$

where Φ_{n+2} is the approximation of the value function Φ_{n+1}. Intuitively, the results will be more accurate if adopting more steps ahead. It has less computational complexity due to its approximation and truncation rules. However, this adaptation can be transformed to near optimality if it executes $(\mathcal{T}+1)$-step look-ahead policy. The computational complexity of one-step look-ahead method is $\mathcal{O}(k\mathcal{N}^2)$, which is de-

termined only by the buffer size and the number of admissible service rates in the current stage. Since in practice, $\mathcal{N} \ll \mathcal{T}$, we have $\mathcal{O}(k\mathcal{N}^2) \ll \mathcal{O}(k\mathcal{N}\mathcal{T})$.

4.4 Simulation Results

4.4.1 Simulation Setup

In the simulation, the HeNB is postulated to maintain two separate queues for the HP and LP users. Both queues follow the FCFS policy and share the same buffer size k. Following the discussion in [68, 69], the arrival rates of the HP and LP users are set as $\lambda_1 = 0.05$ users/slot and $\lambda_2 = 0.2$ users/slot, respectively. The HP (LP) user arrivals are independent regardless of the server in service or vacation period. The admissible service rates are set as $\mathcal{R} = \{0.2, 0.4, 0.6, 0.8\}$. A slot is normalized to 1 s. The HeNB mode transient period between service and vacation is omitted due to its quite short time period, e.g., 70 μs [35]. Based on the observations from queueing analysis shown in Figs. 4.2 and 4.3, we set the maximum waiting times for the HP and LP users as $\omega_1 = 10$ s and $\omega_1 = 20$ s, respectively, and the thresholds of blocking probabilities for the HP and LP users as $\eta_1 = 0.02$ and $\eta_1 = 0.06$, respectively. All other system parameters are compliant with 3GPP LTE HeNB setup [14]. For all simulation results, we evaluated the system after 1000 slots until the system is in steady state. All the simulations are completed in Matlab R2012b version 8.0.0.783.

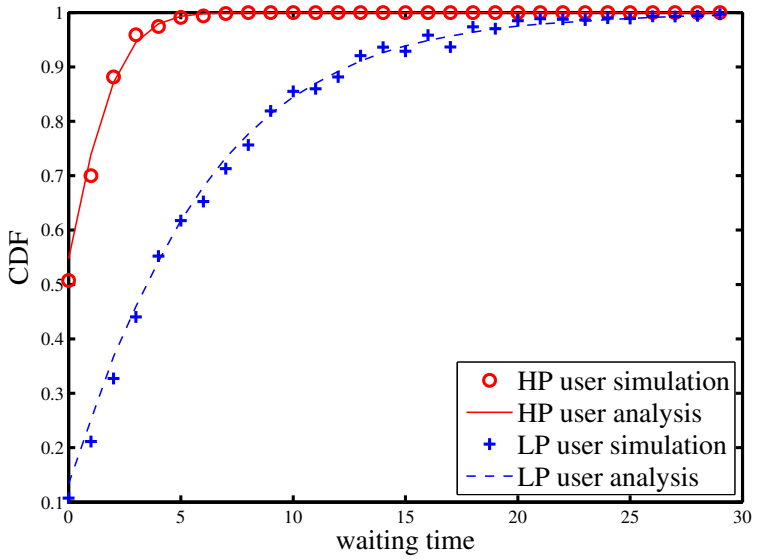


Figure 4.2: The cumulative distribution functions (CDF) of the waiting time (index in slots) for the HP and LP users.

4.4.2 Performance Evaluation

Figs. 4.2 and 4.3 show the key characteristics of the analytical MAP/PH/1/k queue with multiple vacations and user priorities. Fig. 4.2 presents both simulation and analytical results on the cumulative distribution functions (CDF) of the waiting time for both HP and LP users. It is demonstrated that analytical results match the numerical results very well, which validates the accuracy of the proposed analytical queueing model. In addition, it is obvious that the HP user with the privilege of vacation interruption has lower waiting time compared to the LP user. Fig. 4.3 shows the effect of user priorities on blocking probabilities with a variation on vacation length and service rate. Clearly, long vacations can result in high blocking probability. Another observation is that the blocking rate of the HP users is quite lower compared to the LP users. It is because i) when an HP user enters the queue, it always has

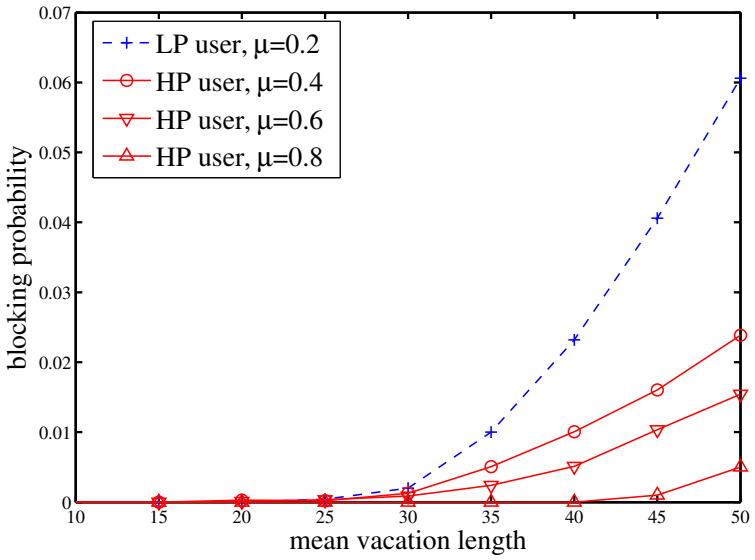


Figure 4.3: Blocking probabilities versus mean vacation length with buffer size $k = 10$.

priority to be served, and *ii*) the HP users can interrupt the vacation periods without waiting in the queue. Note that when a large vacation length is applied, poor QoS performances such as user waiting time and blocking rate become inevitable even though high service rates are selected.

Fig. 4.4 compares the proposed ASV method with a traditional method, in which the HeNB adopts fixed service rate and vacation length. For comparison purpose, the service rates μ_1 and μ_2 are set as the same values, i.e., $\mu_1 = \mu_2$, so that we can highlight the effect of dynamic vacation length on system performance. From the figure, we can observe that for a buffer size $k = 20$, the proposed ASV method outperforms the traditional method in system energy efficiency by up to 20%. Such gain can even be enhanced for larger buffer sizes. In addition, as traffic intensity increases, we observe that the performance improvement due to adaptive vacation length reduces gradually. This is because the server becomes busier with fewer slots arranged for vacations.

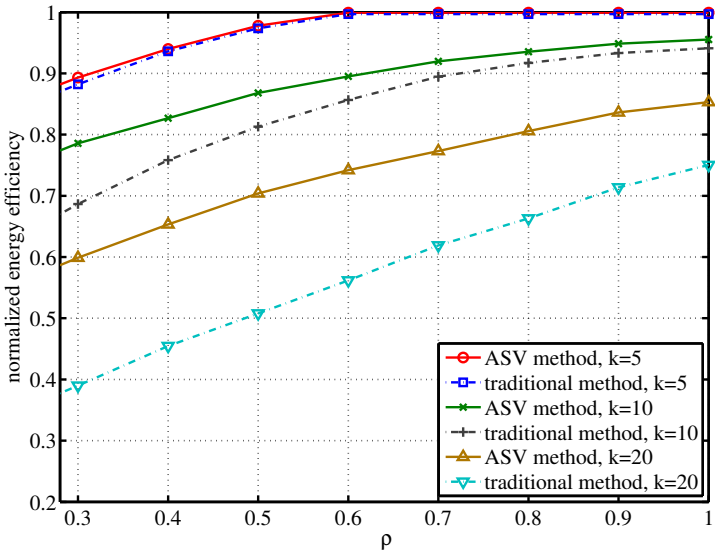


Figure 4.4: The effect of traffic intensity ρ versus system energy efficiency subject to vacation length v and buffer size k.

Such observation suggests that the system should increase buffer size or apply high service rate so as to attain the opportunity of long vacations. In addition, from Fig. 4.4, we can observe that the system energy efficiency degrades with the increase of buffer size because a larger buffer size may result in long and multiple consecutive vacations. Thus, the service period in one busy cycle may become much shorter than the vacation period. Since the vacation and listening periods also consume energy, long and multiple consecutive vacations may reduce energy efficiency due to increased total energy consumption in a busy cycle while keeping system throughput unchanged. In addition to the degradation of energy efficiency, long and multiple vacations may result in the growth of user waiting time. However, the proposed ASV method can make such degradation gradually compared to the traditional method. From an energy saving point of view, the improved energy efficiency implies the HeNB

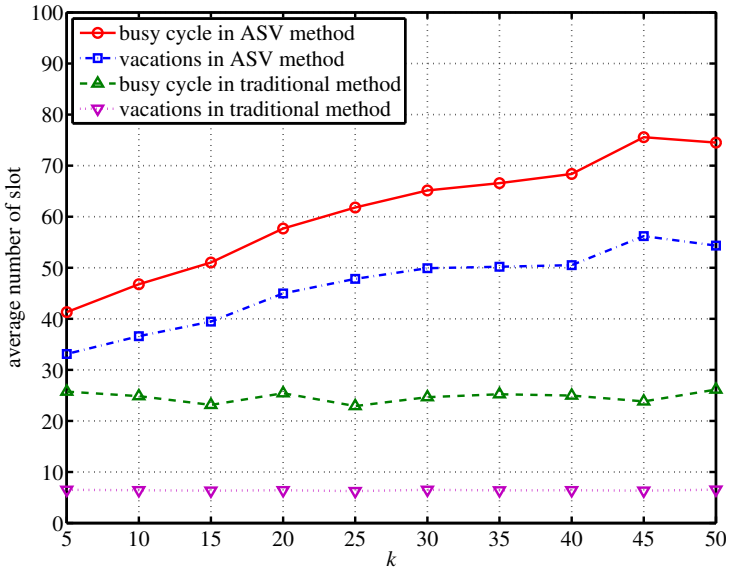


Figure 4.5: Average number of slots per busy cycle and vacation with changing buffer.

experiences more power off periods by suitably adjusting the system parameters. Thus, it can be expected that the proposed ASV method can be more aggressive in performance improvement when the network traffic is low such that the HeNB has more slots for vacations.

Fig. 4.5 compares the average number of busy cycles and vacations between the ASV and the traditional methods. It is obvious that the lengths of busy cycles and vacations change dynamically with the variation of buffer size in the ASV method, while in the traditional method, both performance measures almost remain the same regardless of what buffer size the system adopts. Therefore, the ASV method is more flexible and can better adapt to the dynamics of system parameters.

In Figs. 4.6 and 4.7, the average waiting times and blocking rates for both types of users are presented. The waiting times of the HP users are almost same in both the ASV and the traditional methods, which are close to 0, due to their higher priority

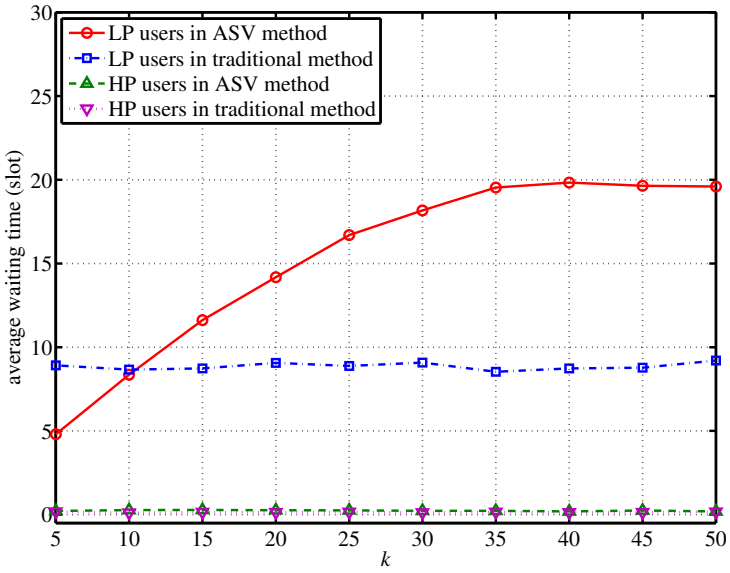


Figure 4.6: Average waiting time for the HP and LP users.

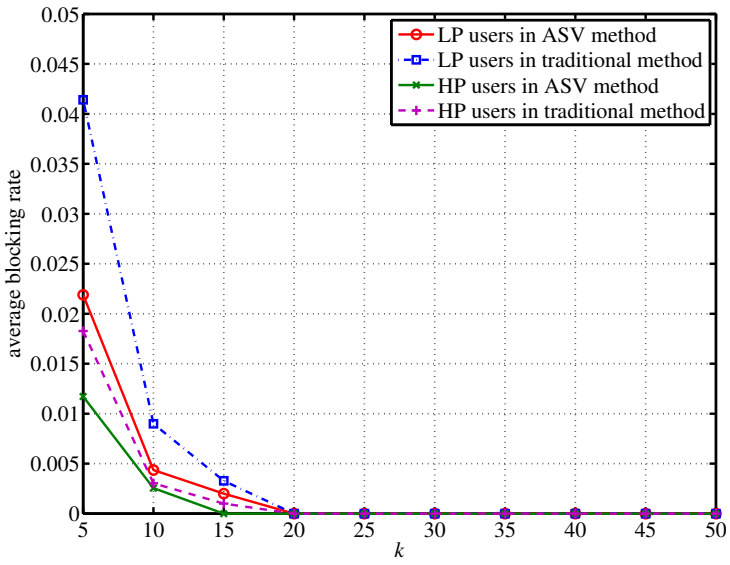


Figure 4.7: Average blocking rate for the HP and LP users.

in service and vacation interruption. However, for the LP users, the average waiting time tends to flatten in the traditional method, while, in the ASV method, the average waiting time may rise as the buffer size increases until it meets the delay threshold.

In Fig. 4.7, it is obvious that the ASV method is better than the traditional method in terms of average blocking rates for both types of users when the HeNB has a small buffer.

Chapter 5

Conclusions and Future Work

5.1 Conclusions

In this research, the energy-efficient strategies with the BS power management for green networking and its related issues in small cell networks have been studied.

First, the BS sleep mode techniques and dynamic BS clustering strategy have been discussed. We addressed the issue that the traditional unified BS-off method is not able to transform effectively with the traffic fluctuations in time and space. Two energy-efficient algorithms: the centralized CBSO protocol with the aid of BSC and the distributed CBSO protocol with the collaboration of BSs have been proposed. Both of them can adaptively adjust and form BS clusters, which can run the optimal BS-off matching scheme separately based on the traffic distribution so as to switch off the underutilized BSs as many as possible. The simulation results have demonstrated that proposed methods have the potential to save more than 50% power consumption and 15.1% more energy efficient compared to the unified BS-off method.

Second, to design a better distributed BS power management strategy for small cell networks, the concept of BS virtual small networks (VSNs) has been proposed. The self-organized VSNs can be formed based on the heterogeneous traffic distribution and the underutilized BSs can be switched off during a certain period. In the proposed *reception* process, each BS can use the method of fuzzy inference system (FIS) to calculate the realistic traffic level based on the collected information including the number of MUs, the average distance between MUs and BS, and the timeline. Then, in the *analysis* process, the method of hidden Markov model (HMM) has been utilized for an individual BS to estimate the mean network traffic based on the collected traffic information from the neighboring BSs. After that, in the *control* process, the BSs VSN formation protocol with firefly (BSVF) algorithm has been proposed, which aims to adaptively group BSs based on the BSs' traffic loads in order to match the heterogeneous traffic variation. Two scenarios with different functions of brightness, which is a traffic load related factor, have been discussed and demonstrated. By jointly considering the traffic load and power consumption in each VSN, the proposed BSVF algorithm has shown the effectiveness on improving load balancing and the potential to use more aggressive BS-off strategies by using the modified BS-off matching scheme. Moreover, the BS power on/off weight has been proposed to prevent frequent BS working mode transitions. The simulation results have demonstrated that the proposed strategies can improve the power saving and load balancing in wireless networks significantly.

Finally, to improve the energy-efficiency for femtocell BS, the detailed operation procedure of the HeNB has been studied. The characteristics of the HeNB with hybrid

access and transmitter power on/off are considered. Accordingly, an MAP/PH/1/k queueing model with user priorities and multiple vacations to better match the HeNB behavior has been discussed. In addition, a vacation termination policy has been designed for benefiting users with a high priority. Based on the formulated queueing model, the key performance metrics in terms of users waiting time, blocking rate and system busy cycle have been derived. Based on the analytical results, a strategy called adaptive service rate and vacation length (ASV) method has been proposed in order to maximize the HeNB's energy efficiency and guarantee QoS provisioning in the same time. Moreover, the one-step look-ahead method has been discussed to reduce the computational complexity. The simulation results have demonstrated that the proposed ASV method can significantly enhance the HeNB's energy efficiency.

5.2 Future Work

Our future work will keep focusing on the development of green wireless networks such as investigating the energy-efficient traffic offloading in heterogeneous networks (HetNets), extending the network lifetime by considering the placement of mobile BSs, and cooperating cellular networks and smart grids via game theory. The snapshots of our future works are provided in the following sections.

5.2.1 Energy-harvesting and Traffic Offloading in HetNets

Currently, the HetNets are promising as an advanced telecommunication network consisting of infrastructure nodes with various wireless access technologies. The HetNets may involve traditional deployment of macrocells with relay stations and many

types of small cells such as microcell, picocells, and femtocell, etc. Each type of them has differed in functionalities, coverages, and constraints. For newly joined small cells in the HetNets, it is important that they can be deployed with a relatively low network overhead, and require little upfront planning and less cost, so as to drastically reducing the operational and capital expenditures of overall wireless networks [70].

In practice, the deployment of small cells coexisting with underlaid macrocells targets to enhance user capacity, improve the indoor coverage and the performance for cell edge users, boost spectral efficiency per unit area via spatial reuse of spectrum, and alleviate the burden of overloaded macrocells [30]. However, such development must face the increase on energy consumption and interference issue due to congested wireless networks. In HetNets, the rising problem is how to cooperate the small cells with the existing macrocells in load balancing and interference coordination, so as to improve the network performance and save power consumption.

We can consider the BS with energy harvesting technique. The BS power consumption consists of two sources: non-renewable energy source and energy harvesting source. The BSs are equipped with the rechargeable batteries of finite capacity and the energy replenishment process can be modeled as a Bernoulli process. For the purpose of saving non-renewable energy, the BSs always adopt the harvesting energy as the first choice to serve users. When the BS runs out of battery, it can be switched to sleep mode for energy replenishment process and assign its associate users to the active BSs for saving non-renewable energy.

By jointly considering the HetNets load balancing, we can define a load biasing factor, ρ, for the BSs, which can be adjusted in the range of $[0, 1]$. Thus, two boundary

cases exist: (i) $\rho_i = 0$ states that there is no load biasing or offloading for the BSs, and (ii) $\rho_i = 1$ means full load biasing is adopted and the BSs can offload all their traffic to the BSs from other tiers. Then, the BSs without load can be switched off.

In the HetNets, when the working status of a BS is switched between the active and sleep modes alternatively, this change may cause a knock-on effect on network performance from many aspects, such that (i) the network energy consumption may vary accordingly; (ii) the throughput and SINR are certainly changing since a new source of interference appears or disappears to affect nearby BSs; and (iii) the load balance is broken and the user association should be re-arranged. Therefore, to improve the system energy efficiency without sacrificing the network performance, it is imperative to design an optimal BSs operation by adaptively managing the BSs on/off strategy and user association. Such BSs operation should achieve the following objectives by jointly considering the BS energy harvesting, load balancing, and the HetNets performance.

In the future work, we can focus on the tradeoff among three essential objectives in the HetNets.

- O_1) minimize the total non-renewable energy consumption.
- O_2) maximize the system throughput.
- O_3) maximize the load balance.

It is obvious these objectives are mutually conflicting. For example, the objective O_1 conflicts with objective O_2 since saving the BSs transmit power without constraints result in decreased SINR and data rates. The contradiction between objectives O_1 and

O_2 is because from capitalizing on energy harvesting perspective, the user association highly depends on the BSs energy level in the battery so that the BSs with more energy storage will serve more users. The objective O_2 is contradicting with objective O_3 because the high system throughput requires maximum SINR user association, which may cause severe load imbalance. To address all these objectives together, a multi-objective optimization problem (MOP) can be adopted. The MOP should consist of three optimization problems, and all problems target to optimize the same variables: the active BSs in each tier and the involved association users, which can optimize the operation of BSs on/off selection and load balancing in the HetNets. The weighted Tchebycheff metric method and the ϵ-constraint method can be considered to solve the MOP in our future work.

5.2.2 Mobile BS Placement with Energy Constraints

In some particular applications of wireless networks, such as military battlefield, breaking news, and emergency operations, traditional fixed BS may be neither available nor reliable to guarantee the prerequisite wireless communications [71]. For example, the operators need to quickly increase service capacity in a certain area during a relatively short time in the case of various types of events, such as sports and concerts with large gatherings of people. It is likely that the pressure on the wireless networks of a certain area will increase dramatically, and the existing wireless service may become jammed and blocked. Thus, to improve the capacity and coverage of wireless networks, deploying mobile BS is a promising solution for a temporary operation in the event area.

A mobile BS has similar functions and pieces of equipment included in an ordinary BS. One difference is that mobile BS can be moved around and set up in a short period. When the fixed BS has been put out of commission for some reason, the mobile BS can rapidly provide coverage. The other typical feature of a mobile BS is that both electricity and cooling must be self-supporting, which means the mobile BS can be deployed in the area without power supply. These features of mobile BS are particularly beneficial to the military application. However, in any emergency or military communication, the battery limitation and off-grid is the biggest challenge for deploying the mobile BSs.

Some studies have considered the BS placement algorithms in wireless networks [72] [73]. In [72], the authors first formulated the BS placement as an integer linear program (ILP) problem. It was assumed the existence of a set of feasible locations, otherwise, the search space for the ILP problem would be infinite. The lifetime of wireless network was divided into equal periods of time called rounds. In each round, the feasible location of a BS was determined in terms of minimum energy consumption of all the BSs in the network. In [73], the mobile BS placement algorithm was proposed based on the boundary of the BS transmission range and the maximum overlap regions with adjacent BSs. It presented that a network lifetime could be improved even if an optimally placed fixed sink (similar to a BS controller), was replaced by a randomly moving BS. However, the consideration of battery limitation of the mobile BS and the load balancing are still missing in the current literature.

In the future work, in order to extend the network lifetime and minimize the power consumption of the mobile BSs, we can consider following energy constraints and load

balancing when placing the mobile BSs.

- The residual energy in a mobile BS must be sufficient to complete the service for the associated mobile users (MUs).
- All the MUs in the system must be covered by the mobile BSs, and when an MU is connected to a mobile BS, the MU must be located in the coverage of the mobile BS.
- The number of the associated MUs with any mobile BS should be larger than a predefined threshold of load balancing, which can avoid the situation that a mobile BS only serves a very small number of users.

Such optimization problem can be considered to solve by computer based simulation with the exhaustive search for the feasible locations of deploying the mobile BSs. However, how to decide the best locations for load balancing and reduce the computation complexity need to be further investigated.

5.2.3 Cooperation of Wireless Networks and Smart Grids

Although some research has been done towards energy-efficient wireless networks, most of them do not consider effects of the power grid, which can provide electricity to the entire networks. Currently, the power grid is experiencing a significant change from the conventional grid to the smart one, which can optimize electricity generation, transmission, and distribution, prevent power blackout, and reduce peaks in power usage, by incorporating computer and information technologies with intelligent control algorithms [74].

By jointly considering the smart grid and wireless networks, the traditional energy-efficient strategies may not be sufficient. In fact, when a system is powered by the smart grid, in some situations, consuming more energy can be better than saving energy. It is because large amounts of highly intermittent and uncontrollable renewable energy (e.g., solar and wind) are integrated with the smart grid. Thus, the cost of energy storage and waste must be considered. When the electricity demand is low, the shutdown of a power plant can be costly or sometimes not technically viable [75]. In addition, the electricity storage capacities are limited in practice [76]. Consequently, when there is redundant electricity in the smart grid, negative prices may occur (e.g., consumers will be paid to encourage to consume more energy) [75] [76]. Therefore, the dynamic operation and management of the smart grid will have significant impacts on improving the energy efficiency of wireless networks.

In [77], it introduced a concept called demand-side management (DSM), which is a set of programs implemented by utility companies that granted customers a greater role in dynamically changing or shifting power consumption. DSM is an important mechanism for helping utilities to operate more efficiently, reduce CO_2-e, and decrease the cost for electricity consumers. If the wireless networks can perform DSM dynamically, the real-time electricity pricing provided by multiple retailers can be considered in the wireless networks [78] [79].

In future work, we can consider a cognitive HetNets with macro-femtocell (outdoor and indoor) configuration, which can be powered by the smart grid. There are multiple electricity retailers in the smart grids and they can provide power to both macro BS (MBS) and femto BS, i.e., HeNBs, with different electricity prices. The

MBS and HeNBs can sense the smart grid, acquire the electricity price, and adjust the amount of electricity they consume by performing energy-efficient power strategies.

In the HetNets scenario, there exist an MBS and multiple HeNBs. All the HeNBs are under the coverage of the MBS over a broadband connection, such as cable modem or digital subscriber line. The MBS can serve macro users in its coverage and offer the unused spectrum to the HeNBs with a price. On the other hand, the HeNB can provide service to the femto users. By utilizing the cognitive radio technology, the MBS is aware of spectrum access by the HeNBs, and the HeNBs can monitor the surrounding radio spectrum environment and are allowed to randomly access the spectrum. The HeNBs will adaptively adjust their power usage or change their access of subbands based on the channel condition and the spectrum price offered by the MBS. It is inevitable that the interference is incurred and the QoS may have degraded since the MBS and HeNBs share the same spectrum resources. By considering the fairness, the interference between MBS and each HeNB can be reflected in the spectrum price. However, the interference among HeNBs can be ignored because the HeNBs are commonly sparsely deployed in the indoor environment.

Then, such a problem can be described as following, which jointly considers the utility maximization of the retailers in the smart grid, and the MBS with HeNBs in the HetNets. First, each retailer in the smart grid offers a real-time electricity price to the MBS and HeNBs. Second, the MBS need to decide which retailer it acquires electricity from, and the amount of the electricity consumed based on the electricity prices. Then, the MBS offers a spectrum price to the HeNBs by considering the channel conditions. Finally, each HeNB decides which retailer to be selected and the

amount of the electricity it will consume based on the electricity prices and whether to use the spectrum offered by the MBS. In this game, each retailer aims to gain as much profit as possible, and the MBS and HeNBs target to spend less and maximize the utilization of resources. We can model the above competition scenario as non-cooperative games with multiple agents, i.e., the first game among retailers, the MBS and HeNBs with the tradeoff on electricity price, and the second game between the MBS and HeNBs with the tradeoff on spectrum price. The detailed analysis and proper algorithms can be developed to reach Pareto optimal for the multiple agents.

Bibliography

[1] M. Ismail and W. Zhuang, *Cooperative networking in a heterogeneous wireless medium*. New York: USA: Springer-Verlag, Apr. 2013.

[2] D. Cavalcanti, D. P. Agrawal, C. Cordeiro, B. Xie, and A. Kumar, "Issues in integrating cellular networks wlans, and manets: a futuristic heterogeneous wireless network," *IEEE Wireless Commun. Mag.*, vol. 12, no. 3, pp. 30–41, Jun. 2005.

[3] G. Fettweis and E. Zimmermann, "ICT energy consumption-trends and challenges," in *Proc. of 11th International Symposium on Wireless Personal Multimedia Communications*, Sep. 2008, pp. 1–6.

[4] *Green radio - NEC's approach towards energy-efficient radio access networks*, NEC Corporation, Tech. Rep., Feb. 2010.

[5] H. Bogucka and A. Conti, "Degrees of freedom for energy savings in practical adaptive wireless systems," *IEEE Commun. Mag.*, vol. 49, no. 6, pp. 38–45, Jun. 2011.

[6] S. Mclaughlin *et al.*, "Techniques for improving cellular radio base station energy efficiency," *IEEE Wireless Commun.*, vol. 18, no. 5, pp. 10–17, Oct. 2011.

[7] *SMART 2020: Enabling the low carbon economy in the information age*, The Climate Group and Global e-Sustainability Initiative (GeSI), 2008, [Online]. Available: http://www.smart2020.org.

[8] Z. Hasan, H. Boostanimehr, and V. K. Bhargava, "Green cellular networks: A survey, some research issues and challenges," *IEEE Communications Surveys & Tutorials*, vol. 13, no. 4, pp. 524–540, Nov. 2011.

[9] A. Amanna, *Green Communications*, Wireless@Virginia Tech., Tech. Rep., Feb. 2010.

[10] A. P. Bianzino, C. Chaudet, D. Rossi, and J.-L. Rougier, "A survey of green networking research," *IEEE Communications Surveys & Tutorials*, vol. 14, no. 1, pp. 3–20, Feb. 2012.

[11] C. Han, T. Harrold, S. Armour, I. Krikidis, S. Videv, P. M. Grant, H. Haas, J. S. Thompson, I. Ku, C.-X. Wang *et al.*, "Green radio: radio techniques to enable energy-efficient wireless networks," *IEEE Commun. Mag.*, vol. 49, no. 6, pp. 46–54, Jun. 2011.

[12] K. J. Christensen, C. Gunaratne, B. Nordman, and A. D. George, "The next frontier for communications networks: power management," *Computer Communications*, vol. 27, no. 18, pp. 1758–1770, Dec. 2004.

[13] M. A. Marsan, L. Chiaraviglio, D. Ciullo, and M. Meo, "Optimal energy savings in cellular access networks," in *Proc. of IEEE International Conference on Communications Workshops (ICC'09, GreenCom Wksp.)*, Jun. 2009, pp. 1–5.

[14] 3GPP TS 36.141 V11.2.0, *Technical Specification Group Radio Access Network, E-UTRA Base Station Conformance Testing, Release 11*, Nov. 2012.

[15] E. Oh, B. Krishnamachari, X. Liu, and Z. Niu, "Toward dynamic energy-efficient operation of cellular network infrastructure," *IEEE Commun. Mag.*, vol. 49, no. 6, pp. 56–61, Jun. 2011.

[16] Z. Niu, Y. Wu, J. Gong, and Z. Yang, "Cell zooming for cost-efficient green cellular networks," *IEEE Commun. Mag.*, vol. 48, no. 11, pp. 74–79, Nov. 2010.

[17] X. Weng, D. Cao, and Z. Niu, "Energy-efficient cellular network planning under insufficient cell zooming," in *Proc. of IEEE 73rd Vehicular Technology Conference (VTC-Spring'11 Greenet Wksp.)*, May 2011, pp. 1–5.

[18] S. Bhaumik, G. Narlikar, S. Chattopadhyay, and S. Kanugovi, "Breathe to stay cool: adjusting cell sizes to reduce energy consumption," in *Proc. of 1st ACM SIGCOMM Workshops on Green Networking*, Sep. 2010, pp. 41–46.

[19] W. T. Wong, Y. J. Yu, and A. C. Pang, "Decentralized energy-efficient base station operation for green cellular networks," in *Proc. of IEEE Global Communications Conference (GLOBECOM'12)*, Dec. 2012, pp. 5194–5200.

[20] R. Wang, J. S. Thompson, H. Haas, and P. M. Grant, "Sleep mode design for green base stations," *IET Commun.*, vol. 5, no. 18, pp. 2606–2616, Jan. 2011.

[21] A. Bousia, A. Antonopoulos, L. Alonso, and C. Verikoukis, "'green distance-aware base station sleeping algorithm in lte-advanced," in *Proc. of IEEE ICC'12*, Jun. 2012, pp. 1347–1351.

[22] F. Han, Z. Safar, W. S. Lin, Y. Chen, and K. J. Liu, "Energy-efficient cellular network operation via base station cooperation," in *Proc. of IEEE ICC'12*, Jun. 2012, pp. 4374–4378.

[23] J. Hoydis, M. Kobayashi, and M. Debbah, "Green small-cell networks," *IEEE Veh. Technol. Mag.*, vol. 6, no. 1, pp. 37–43, Mar. 2011.

[24] H. Leem, S. Baek, and D. Sung, "The effects of cell size on energy saving, system capacity and per-energy capacity," in *Proc. of IEEE Wireless Communications and Networking Conference (WCNC)*, Apr. 2010, pp. 1–6.

[25] G. De La Roche, A. Valcarce, D. López-Pérez, and J. Zhang, "Access control mechanisms for femtocells," *IEEE Commun. Mag.*, vol. 48, no. 1, pp. 33–39, Jan. 2010.

[26] Y. Chen, S. Zhang, S. Xu, and G. Y. Li, "Fundamental trade-offs on green wireless networks," *IEEE Commun. Mag.*, vol. 49, no. 6, pp. 30–37, Jun. 2011.

[27] *LTE Advanced: Heterogeneous Networks*, Qualcomm Inc., Tech. Rep., Jan. 2011, [Online]. Available: http://www.qualcomm.com/media/documents/lte-heterogeneous-networks.

[28] C. Prehofer and C. Bettstetter, "Self-organization in communication networks: principles and design paradigms," *IEEE Commun. Mag.*, vol. 43, no. 7, pp. 78–85, Jul. 2005.

[29] H. ElSawy, E. Hossain, and D. I. Kim, "Hetnets with cognitive small cells: User offloading and distributed channel access techniques," *IEEE Commun. Mag.*, vol. 51, no. 6, Jun. 2013.

[30] D. Lopez-Perez, I. Guvenc, G. De La Roche, M. Kountouris, T. Q. Quek, and J. Zhang, "Enhanced intercell interference coordination challenges in heterogeneous networks," *IEEE Wireless Commun.*, vol. 18, no. 3, pp. 22–30, Jun. 2011.

[31] O. Aliu, A. Imran, M. Imran, and B. Evans, "A survey of self organisation in future cellular networks," *IEEE Communications Surveys & Tutorials*, vol. 15, no. 1, pp. 336–361, Feb. 2012.

[32] J. G. Andrews, H. Claussen, M. Dohler, S. Rangan, and M. C. Reed, "Femtocells: Past, present, and future," *IEEE J. Sel. Areas Commun.*, vol. 30, no. 3, pp. 497–508, Sep. 2012.

[33] 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 Trans. Wireless Commun.*, vol. 12, no. 6, pp. 2706–2716, Apr. 2013.

[34] A. Barbieri *et al.*, "Lte femtocells: system design and performance analysis," *IEEE J. Sel. Areas Commun.*, vol. 30, no. 3, pp. 586–594, Apr. 2012.

[35] 3GPP TR 36.921, *Evolved Universal Terrestrial Radio Access (E-UTRA); FDD Home eNode B (HeNB) Radio Frequency (RF) requirements analysis, Release 9*, Mar. 2010.

[36] L. Saker, S. E. Elayoubi, R. Combes, and T. Chahed, "Optimal control of wake up mechanisms of femtocells in heterogeneous networks," *IEEE J. Sel. Areas Commun.*, vol. 30, no. 3, pp. 664–672, Apr. 2012.

[37] W. Vereecken, M. Deruyck, D. Colle, W. Joseph, M. Pickavet, L. Martens, and P. Demeester, "Evaluation of the potential for energy saving in macrocell and femtocell networks using a heuristic introducing sleep modes in base stations," *EURASIP J. Wireless Commun. Net.*, vol. 2012, no. 170, pp. 1–14, May 2012.

[38] R. Estrada, A. Jarray, H. Otrok, Z. Dziong, and H. Barada, "Energy-efficient resource-allocation model for ofdma macrocell/femtocell networks," *IEEE Trans. Veh. Technol.*, vol. 62, no. 7, pp. 3429–3437, Sep. 2013.

[39] Y. Li, H. Celebi, M. Daneshmand, C. Wang, and W. Zhao, "Energy-efficient femtocell networks: challenges and opportunities," *IEEE Wireless Commun.*, vol. 20, no. 6, pp. 99–105, Dec. 2013.

[40] X. Ge, T. Han, Y. Zhang, G. Mao, C. Wang, J. Zhang, B. Yang, and S. Pan, "Spectrum and energy efficiency evaluation of two-tier femtocell networks with partially open channels," *IEEE Trans. Veh. Technol.*, vol. 63, no. 3, pp. 1306–1319, Mar. 2014.

[41] J. Kim, W. S. Jeon, and D. G. Jeong, "Effect of base station-sleeping ratio on energy efficiency in densely deployed femtocell networks," *IEEE Commun. Lett.*, vol. 19, no. 4, pp. 641–644, Apr. 2015.

[42] A. S. Alfa, "Matrix-geometric solution of discrete time map/ph/1 priority queue," *Naval Research Logistics*, vol. 45, no. 1, pp. 23–50, Feb. 1998.

[43] A. S.Alfa, *Queueing theory for telecommunications: discrete time modelling of a single node system*. Springer Science & Business Media, 2010.

[44] A. S. Alfa, B. Liu, and Q. M. He, "Discrete-time analysis of map/ph/1 multiclass general preemptive priority queue," *Naval Research Logistics*, vol. 50, no. 6, pp. 662–682, Sep. 2003.

[45] L. Chiaraviglio, D. Ciullo, M. Meo, M. A. Marsan, and I. Torino, "Energy-aware UMTS access networks," in *Proc. of 11th International Symposium on Wireless Personal Multimedia Communications (W-GREEN'08)*, Sep. 2008, pp. 1–8.

[46] O. Arnold, F. Richter, G. Fettweis, and O. Blume, "Power consumption modeling of different base station types in heterogeneous cellular networks," in *Proc. of 19th Future Network and Mobile Summit*, Jun. 2010, pp. 1–8.

[47] M. F. Hossain, K. S. Munasinghe, and A. Jamalipour, "A protocooperation-based sleep-wake architecture for next generation green cellular access networks," in *4th International Conference on Signal Processing and Communication Systems (ICSPCS)*, Dec. 2010, pp. 1–5.

[48] X. Weng, D. Cao, and Z. Niu, "Energy-efficient cellular network planning under insufficient cell zooming," in *Proc. of IEEE 73rd Vehicular Technology Conference (VTC Spring'11)*, May 2011, pp. 1–5.

[49] *ITU-R, Guidelines for evaluation of radio interface technologies for IMT-Advanced*, Rep. ITU-R M.2135-1, Geneva, Switzerland, Dec. 2009.

[50] M. F. Hossain, K. S. Munasinghe, and A. Jamalipour, "A protocooperation-based sleep-wake architecture for next generation green cellular access networks," in *Proc. of 4th International Conference on Signal Processing and Communication Systems (ICSPCS'10)*, Dec. 2010, pp. 1–8.

[51] F. Richter, A. J. Fehske, and G. P. Fettweis, "Energy efficiency aspects of base station deployment strategies for cellular networks," in *Proc. of 70th Vehicular Technology Conference Fall (VTC-Fall'09)*, Sep. 2009, pp. 1–5.

[52] H. Shu, Q. Liang, and J. Gao, "Wireless sensor network lifetime analysis using interval type-2 fuzzy logic systems," *IEEE Trans. Fuzzy Syst.*, vol. 16, no. 2, pp. 416–427, Apr. 2008.

[53] J. S. Lee and W. L. Cheng, "Fuzzy-logic-based clustering approach for wireless sensor networks using energy predication," *IEEE Sensors J.*, vol. 12, no. 9, pp. 2891–2897, Sep. 2012.

[54] L. A. Zadeh, "Fuzzy algorithms," *Information and Control*, vol. 12, no. 2, pp. 94–102, Feb. 1968.

[55] E. Cox, "Fuzzy fundamentals," *IEEE Spectr.*, vol. 29, no. 10, pp. 58–61, Oct. 1992.

[56] D. Burshtein, "Robust parametric modeling of durations in hidden markov models," *IEEE Audio, Speech, Language Process.*, vol. 4, no. 3, pp. 240–242, Aug. 1996.

[57] H. L. Lou, "Implementing the viterbi algorithm," *IEEE Signal Process. Mag.*, vol. 12, no. 5, pp. 42–52, Sep. 1995.

[58] X. S. Yang, *Nature-inspired metaheuristic algorithms*. Luniver Press, 2010.

[59] J. Fang and H. Li, "Power constrained distributed estimation with cluster-based sensor collaboration," *IEEE Trans. Wireless Commun.*, vol. 8, no. 7, pp. 3822–3832, Jul. 2009.

[60] ITU-R, *Guidelines for evaluation of radio interface technologies for IMT-Advanced*, Geneva, Switzerland, Rep. ITU-R M.2135-1, Tech. Rep., Dec. 2009, [Online]. Available: http://www.itu.int/pub/R-REP-M.2135/en.

[61] -, *2012 guidelines to Defra/DECC's GHG conversion factors for company reporting*, Department for Environment, Food and Rural Affairs (Defra), UK, Tech. Rep., Aug. 2012, [Online]. Available: http://www.defra.gov.uk/publications/2012/05/30/pb13773-2012-ghg-conversion/.

[62] W. Vereecken, W. V. Heddeghem, D. Colle, M. Pickavet, and P. Demeester, "Overall ICT footprint and green communication technologies," in *Proc. of 4th International Symposium on Communications, Control and Signal Processing (ISCCSP'10),*, Mar. 2010, pp. 1–6.

[63] D. Astely *et al.*, "Lte release 12 and beyond," *IEEE Commun. Mag.*, vol. 51, no. 7, pp. 154–160, Jul. 2013.

[64] W. K. Grassmann, M. I. Taksar, and D. P. Heyman, "Regenerative analysis and steady state distributions for markov chains," *Operations Research*, vol. 33, no. 5, pp. 1107–1117, Sep. 1985.

[65] D. P. Palomar and M. Chiang, "A tutorial on decomposition methods for network utility maximization," *IEEE J. Sel. Areas Commun.*, vol. 24, no. 8, pp. 1439–1451, Aug. 2006.

[66] I. N. Bronshtein, K. A. Semendiaev, G. Musiol, and H. Muhlig, *Handbook of Mathematics 5th Ed.* Springer, 2007.

[67] C. Watkins and P. Dayan, "Q-learning," *Mach. Learn*, vol. 8.

[68] S. Akbarzadeh, R. Combes, and Z. Altman, "Network capacity enhancement of ofdma system using self-organized femtocell off-load," in *Proc. of IEEE Wireless Communications and Networking Conference (WCNC)*, Apr. 2012, pp. 1234–1238.

[69] L. B. Le, D. Niyato, E. Hossain, D. I. Kim, and D. T. Hoang, "Qos-aware and energy-efficient resource management in ofdma femtocells," *IEEE Trans. Wireless Commun.*, vol. 12, no. 1, pp. 180–194, Jan. 2013.

[70] J. Hoydis, M. Kobayashi, and M. Debbah, "Green small-cell networks," *IEEE Veh. Tech. Mag.*, vol. 6, no. 1, pp. 37–43, Mar. 2011.

[71] R. Sánchez, J. Evans, and G. Minden, "Networking on the battlefield: Challenges in highly dynamic multi-hop wireless networks," in *Proc. of IEEE Military Communications Conference, (MILCOM'99)*, vol. 2, Dec. 1999, pp. 751–755.

[72] S. R. Gandham, M. Dawande, R. Prakash, and S. Venkatesan, "Energy efficient schemes for wireless sensor networks with multiple mobile base stations," in *Proc. of IEEE Global Telecommunications Conference (GLOBECOM'03)*, vol. 1, Dec. 2003, pp. 377–381.

[73] W. Alsalih, S. Akl, and H. Hassanein, "Placement of multiple mobile base stations in wireless sensor networks," in *Proc. of IEEE International Symposium on Signal Processing and Information Technology*, Dec. 2007, pp. 229–233.

[74] X. Yu, C. Cecati, T. Dillon, and M. G. Simoes, "The new frontier of smart grids," *IEEE Ind. Electro. Mag.*, vol. 5, no. 3, pp. 49–63, Sep. 2011.

[75] D. Keles, M. Genoese, D. Möst, and W. Fichtner, "Comparison of extended mean-reversion and time series models for electricity spot price simulation considering negative prices," *Energy Economics*, vol. 34, no. 4, pp. 1012–1032, Jul. 2012.

[76] M. Nicolosi, "Wind power integration and power system flexibility–an empirical analysis of extreme events in germany under the new negative price regime," *Energy Policy*, vol. 38, no. 11, pp. 7257–7268, Nov. 2010.

[77] G. M. Masters, *Renewable and efficient electric power systems*. John Wiley & Sons, 2013.

[78] A. J. Conejo, J. M. Morales, and L. Baringo, "Real-time demand response model," *IEEE Trans. Smart Grid*, vol. 1, no. 3, pp. 236–242, Dec. 2010.

[79] A. H. Mohsenian Rad and A. Leon-Garcia, "Optimal residential load control with price prediction in real-time electricity pricing environments," *IEEE Trans. Smart Grid*, vol. 1, no. 2, pp. 120–133, Sep. 2010.

List of Publications

[1] H. Zhang, J. Cai, X. Li, and S. Huang, "Adaptive service rate and vacation length for energy-efficient HeNB based on queueing analysis," *IEEE Trans, Veh. Tech.*, no. 99, pp. 1–14, Dec. 2015.

[2] S. Huang, J. Cai, H. Chen, and H. Zhang, "Transmit power optimization for amplify-and-forward relay networks with reduced overheads," *IEEE Trans, Veh. Tech.*, no. 99, pp. 1–12, Jul. 2015.

[3] H. Zhang, J. Cai, and X. Li, "Self-organized virtual small networking for energy saving and load balancing in cellular networks," in *Proc. IEEE International Conference on Communication Workshop (ICCW'15)*, Jun. 2015, pp. 2874–2879.

[4] X. Li, M. Peng, J. Cai, C. Yi, and H. Zhang, "OPNET-based modeling and simulation of mobile Zigbee sensor networks," *Peer-to-Peer Networking and Applications*, pp. 1–10, Apr. 2015.

[5] H. Zhang and J. Cai, "Energy efficient strategies with BS sleep mode in green small cell networks," in *Design and Deployment of Small Cell Networks*, A. Anpalagan, M. Bennis, and R. Vannithamby, Eds. Cambridge University Press, 2014, ch. 10, pp. 284–308.

[6] X. Li, J. Cai, H. Chen, and H. Zhang, "TLDTCA: a distributed approach to meeting heterogenous connectivity requirements to sink in M2M networks," in *Proc. IEEE Global Communications Conference (GLOBECOM'13)*, Austin, TX, USA, Dec. 2013, pp. 4459–4464.

[7] H. Zhang, J. Cai, and X. Li, "Energy-efficient base station control with dynamic clustering in cellular network," in *Proc. 8th International Conference on Communications and Networking in China (CHINACOM'13)*, Guilin, China, 2013, pp. 384–388.

[8] X. Li, J. Cai, and H. Zhang, "Topology control for heterogeneous connectivity requirements to sink in M2M networks," in *Proc. 8th International Conference on Communications and Networking in China (CHINACOM'13)*, Guilin, China, Aug. 2013, pp. 575–580, **best paper awards**.