Intelligent On-demand Radio Resource Provisioning for Green Ultra-small Cell Networks

Zhehan Li

Ph.D.

University of York
Electronics

September, 2016

Abstract

This thesis studies intelligent on-demand radio resource provisioning involving sleep mode operation in ultra Small Cell Networks (SCNs). Sleep modes are low power states of base stations. The purpose of the research is to investigate how appropriate traffic information can be adopted in sleep mode operation schemes for SCNs with different architectures.

A novel protocol-friendly sleep mode operation algorithm based on Adaptive Traffic Perception is proposed for distributed SCN architectures. It is proved robust to different SCN layouts with the reduction in the average power consumption of base stations being more than 35% while maintaining the Quality of Service.

The Traffic-aware Cell Management scheme adopting Direction of Arrival information is particularly designed to eliminate the necessity of computation for sleeping base stations. This scheme is shown to significantly reduce the side effects associated with the sleep mode operation, including system overheads and the increasing user transmission power.

For SCNs using centralised architectures, such as Cloud Radio Access Networks, Hotspot-oriented Green Frameworks are proposed for different information availabilities, which achieve almost 80% reduction in power consumption of Remote Radio Heads at low traffic levels. A clustering technique is utilised for the optimisation of the placement of active Remote Radio Heads, lowering the average user transmission power. The amount of reduction depends on the completeness of the information and can exceed 70% compared with the state-of-the-art.

A type II Matérn Hard-core Point Process is used for modelling SCNs. The derivation and approximation of its distance distributions are also proposed. The distance distributions are used for the probabilistic theoretical analysis of some metrics of the sleep mode operation.

Contents

List of Figures

List of Tables

Acknowledgments

First, my heartfelt and foremost gratitude goes to my first supervisor, Prof. David Grace. He has been supporting my research by providing lots of helpful advices since my first day as a research student. He led me to the hottest research topics and offered me great freedom to choose my research directions. My research work would not be that successful without his illuminating guidance.

I would also like to thank my second supervisor, Dr. Paul Mitchell and my thesis advisor, Mr. Tim Clarke. They gave me many useful suggestions and assistance on my research life.

My thanks also go to all of my former and present colleagues in the Communications Research Group. They together built a comfortable, friendly and helpful atmosphere for my research life.

Last but not least, I would like to express my sincere and deepest gratitude to my parents, who have been giving their enormous support and selfless love.

Declaration

This work has not previously been presented for an award at this, or any other, University. All the work presented in this report as original is so, to the best knowledge of the author. References and acknowledgements to other researchers have been given as appropriate.

Elements of the research presented in this thesis have resulted in publications and submissions to journals or conferences. A list is provided below.

Journal Papers

1. Z. Li, D. Grace, and P. Mitchell, "Traffic-aware cell management for green ultra-dense small cell networks," *IEEE Trans. Veh. Technol.* [Online]. Available: http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=7484757&isnumber=4356907

2. Z. Li, D. Grace, and P. Mitchell, "On the distance distributions of type II Matérn hard-core cellular networks," *IET Electron. Lett.* [Under review]

3. Z. Li, D. Grace, and P. Mitchell, "Hotspot-oriented green frameworks for ultra-small cell cloud radio access networks," *IEEE Trans. Veh. Technol.* [Under review]

Conference Papers

1. Z. Li, D. Grace, and P. Mitchell, "Traffic perception based topology management for 5G green ultra-small cell networks," in *Proc. IEEE 1st Int. Workshop Cognitive Cellular Syst.*, Sep. 2014, pp. 1-5

2. Z. Li, D. Grace, and P. Mitchell, "Cell division, migration and death for energy efficient 5G ultra-small cell networks," in *Proc. IEEE Globecom Workshops,* Dec. 2014, pp. 942-947

Chapter 1

Introduction

Contents

1.1 Motivation

With the advent of the mobile Internet age and the proliferation of the mobile device market, the past decade has witnessed a highly increasing mobile data demand generated by the explosive growth of emerging Internet services. As recently reported in [8], mobile data traffic in 2013 is nearly 18 times the size of the entire global internet in 2000 and a compound annual growth rate of 61 percent is predicted, which will raise the mobile data traffic to 15.9 exabytes per month by 2018. It is obvious that the dramatically increasing demand for mobile data services imposes great pressure on the capacity of existing wireless networks, which mainly consist of traditional cellular Base Stations (BSs). To keep pace with the market demand, it is essential to provide more radio resources per unit area, i.e. boost capacity densities, especially in urban areas, where high data traffic densities are expected.

As spectral efficiency of a point-to-point link in cellular networks approaches its theoretical limit [2], other techniques such as massive Multiple-input Multiple-output (MIMO) and millimetre Wave (mmW) have attracted more attention of researchers in recent years. The other promising solution is the deployment of dense Small Cell Networks (SCNs) in areas of high data traffic densities, which is well known as the simplest and the most effective way

to enhance system capacity [9]. In SCNs, cell sizes are reduced to improve spatial frequency reuse, meaning that there are higher densities of BSs of smaller sizes and lower cost, supplying more spectrum resources in a certain area.

Although it remains unknown whether traditional macro cell networks will be fully substituted by dense SCNs, the progressive development of light BSs has led to the coexistence of different types of BSs, yielding Heterogeneous and Small cell Networks (HetSNets). In this regard, traditional macro cell networks still exist to guarantee the coverage while SCNs undertake the responsibility of offloading the data traffic from macro cell BSs, providing higher data rates and dedicated capacity to residential areas or other hotspot areas [10]. As a result of the expanding scale of BS deployment, the most worrying issue is the significant increase in the overall power consumption brought by the operation of additional small cell BSs.

With the deployment of more infrastructure, the increasing electricity bill becomes a large burden in controlling the Operational Expenditure (OPEX), which pressurises operators on profitability. Motivated by this reason, small cell BSs are required to have low operation power and low transmit power to reduce the overall power consumption. Beyond this, a low power sleep mode with a limited functionality can also be designed to further reduce power consumption of a BS. As there is a fluctuation of the traffic demand over space and time [11], the overall network power consumption is expected to be reduced by on-demand provisioning of radio resources where some BSs in a network are appropriately switched to sleep modes if necessary. Where and when the radio resources should be supplied has to be determined by proper adaptive schemes to accommodate the varying traffic.

1.2 Hypothesis

The hypothesis guiding the research presented in this thesis is as follows:

> *Appropriate traffic information can be exploited in sleep mode operation to improve on-demand radio resource provisioning for ultra-small cell networks of different architectures.*

On-demand radio resource provisioning is defined as providing just enough radio resources for local areas with various service demands. The evaluation of the improvement with the approach of sleep mode operation includes the amount of the reduction in the overall power consumption and the maintenance of Quality of Service (QoS) compared with no application

of sleep modes. The assessment can be extended to additional aspects considering the side effects such as system overhead and user power consumption. The thesis is focused on designing sleep mode operation schemes to effectively quantify traffic information for decision making considering different potential SCN architectures and infrastructure availabilities. Simulation experiments and comprehensive performance analysis have been carried out to prove the hypothesis.

1.3 Objectives

The objectives are determined based on the hypothesis as follows:

- Sleep mode operation for potential architectures of future ultra-small cell networks should be investigated.
- The difference in distributions of ultra-small cell networks should be characterised.
- Appropriate formats of traffic information and various sleep mode functionalities should be considered in the design of sleep mode operation schemes.
- The performance of on-demand radio resource provisioning should be evaluated by simulation.
- Suitable mathematical tools should be developed to support the analysis of relevant metrics of sleep mode operation.

1.4 Thesis Outline

The rest of the thesis is organised as follows.

Chapter 2 provides a literature review about the background and the established work related to the thesis. The main technologies for the next generation high capacity wireless networks are discussed, where SCNs as the solution studied in the thesis are reviewed. The spatial modelling techniques are introduced for SCNs since they have different distribution patterns from traditional macro BSs. Green technologies for the purpose of reducing the additional power consumption brought by the deployment of SCNs are introduced. A comprehensive review is provided for energy efficient on-demand radio resource provisioning

involving sleep mode operation, which is the main topic of the thesis. The details of sleep modes and the categorisation of their controlling schemes are presented.

Chapter 3 introduces the methodology of the research, including the system models and the performance evaluation methods. The procedures of the simulation based work are described.

Chapter 4 studies a stochastic model, type II Matérn Hard-core Point Process (MHPP), with its parent point process, the Homogeneous Poisson Point Process (HPPP) introduced as the fundamental. The model is studied for modelling the distinct spatial distribution of BSs in SCNs and is adopted in the simulation presented in the thesis. The distance distributions are investigated, where they are derived and approximated for MHPP. The approximation results form the basis of the theoretical analysis of some related metrics in sleep mode operation in the following chapters.

Chapter 5 proposes the Adaptive Traffic Perception (ATP) based sleep mode operation algorithm for distributed SCNs. A concept of adaptive monitoring areas is introduced. The numbers of active links in the monitoring areas are monitored. They as well as the sizes of monitoring areas are exploited as the traffic information. The practical implementation of the algorithm is described. Its performance is evaluated on a grid based model and a comparison with the MHPP model is provided. It is also theoretically analysed on the MHPP model and the relationship among some metrics of the ATP based sleep mode operation are correlated with an equation.

Chapter 6 proposes a Traffic-aware Cell Management scheme to improve the ATP based sleep mode operation algorithm in terms of power consumption of sleep modes, UE transmission power and system overhead. It utilises the Direction of Arrival (DOA) information to form the awareness of the direction to hotspot areas for active BSs in a distributed SCN. The quantified awareness and the loads of BSs are the traffic information exploited. The performance of the scheme is evaluated on the SCNs modelled by MHPPs and compared with a recent consolidated scheme.

Chapter 7 studies sleep mode operation in a centralised architecture, Cloud Radio Access Network (C-RAN), where different frameworks involving clustering techniques are proposed to control the state transitions of Remote Radio Heads (RRHs). The clustering techniques are expected to exploit the quantified traffic distribution for the optimisation of active RRH placement. It is for the purpose of reducing the UE transmission power, which is also theoret-

ically analysed with the MHPP model. The quantified traffic information in the frameworks is different for different information availabilities. One of the frameworks is compared with an existing strategy dealing with hotspot areas as the baseline scheme.

Chapter 8 concludes the contributions made by the thesis and provides recommendations on potential extensions of the proposed work.

Chapter 2

Literature Review

Contents

2.1 Introduction

The thesis investigates on-demand radio resource provisioning for ultra-small cell networks as indicated in the hypothesis. Small Cell Networks (SCNs) take the responsibility of capacity enhancement in the future heterogeneous and small cell networks (HetSNets), which are expected as an energy and cost efficient solution to provide ultra-high capacity wireless services. In SCNs, different types of BSs, multiple radio access technologies (Multi-RATs) and several optional architectures may be applied for diverse purposes. The large-scale deployment of SCNs brings a different distribution pattern of BSs, which needs to be investigated and new spatial modelling techniques are required. On the other hand, the overall power consumption is growing rapidly along with the extensive deployment of new SCNs. Green networking therefore becomes a parallel topic, where various technologies have been studied to alleviate power consumption. Among the technology candidates, energy saving potential in BSs and network architectures is expected to be important. The aforementioned on-demand radio resource provisioning and sleep modes also attract much attention because they can achieve a large amount of reduction in network power consumption while delivering ultra-high capacity services.

This chapter presents the related background information and provides a review of the published research about the above topics. An overview of candidate technologies for ultra-high capacity wireless networks is given in Section 2.2, where SCNs are introduced. In Section 2.3, the spatial models of BSs in SCNs are discussed and compared. In Section 2.4, green networking is comprehensively reviewed with different enabling technologies provided and on-demand radio resource provisioning particularly presented. Sleep modes and representative operation algorithms are categorised and reviewed with more detail. Finally, the summary is given in Section 2.5.

2.2 Ultra-high Capacity Wireless Networks

To meet the continuously increasing demand for mobile data services, industry is making effort to enable radio access networks (RANs) to have a further 1000× capacity by 2020 relative to the wireless networks in 2010 [12]. It is worth mentioning that capacity here is in bits per second per square kilometre (bps/km^2), which refers to capacity density sometimes. Rethinking the definition, the capacity unit can be written as bps per Hz per cell × Hz ×

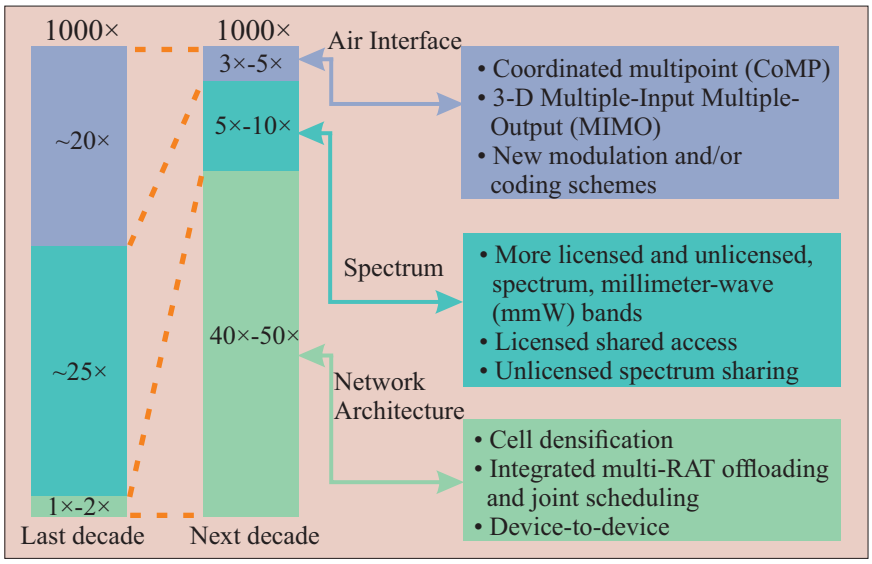


Figure 2.1: 1000× network capacity growth [1].

cell per square kilometre. The three variables indicate three main approaches of capacity improvement, which are air interface, spectrum and network architecture, respectively. Their relationship and developments are categorised and summarised in Fig. 2.1.

In the last decade, air interface improvements and new spectrum allocations have made a significant contribution to capacity upgrades of wireless networks, roughly 20 times and 25 times, respectively. However, as time goes by, spectrum efficiency of the air interface has been approaching its capacity limit [1]. Recent research is more focused on the benefits from the enhancement of "× Hz" and "× cell per square kilometre" to realise the ultra-high capacity target. The main approaches of the capacity enhancement have been shifted to other technologies to form a more accommodative network architecture, which is expected to provide 40-50 times capacity increase. In the vision of the future network architecture, cell densification plays an important role where multiple radio access technologies (Multi-RATs) are integrated to ensure a large number of access nodes and thus sufficient radio resources in a certain area. As the research in the thesis is on SCNs, their future logical architectures in the following decade is investigated and the main challenges are reviewed in this section.

2.2.1 Small Cell Networks

As directly indicated from the term × cell per square kilometre, one of the promising paradigms to boost capacity is to have more cells in a certain area, i.e. to deploy more BSs with enhanced frequency reuse. In the past decades, a traditional macro base station

(MBS) had the coverage range up to a few tens of kilometres, which had a limited amount of frequency resources per unit area. Much recent work has been devoted to overlaying low-power and low-cost nodes elsewhere in a macrocell, serving to offload data traffic from a MBS. While MBSs still exist to provide a blanket coverage and handle high mobility users, these small-sized nodes form Small Cell Network (SCNs) to supply more radio resources to certain hotspot areas and improve capacity. As a result, MBSs and small cell nodes constitute complex networks of different tiers, among which an appropriate management standard is required. On the other hand, considering the co-tier coordination and cooperation among small cell nodes, the single tier architecture of a future SCN has two options: the distributed architecture and the centralised architecture. They are both considered in the research presented in the thesis, so they are reviewed here.

Distributed Small Cell Networks

In the conventional distributed architecture, there are no centralised controllers managing the operation of the co-tier BSs. The intelligence is distributed amongst BSs and the neighbour BSs are interconnected through a defined air interface. The coordination and cooperation of BSs rely on the information exchanged via the air interface. For such a BS, radio frequency (RF) modules, including the power amplifier and the antenna module, are located close to its dedicated baseband unit (BBU) so that each BS can transmit, receive and process signals in its own coverage area [13].

A representative example of the distributed architecture is the current Long Term Evolution-Advanced (LTE-A) system. The LTE standard has been developed by the Third Generation Partnership Project (3GPP) since Release 8, which was published in March, 2009. The LTE standard has been practically realised in the current fourth generation (4G) wireless systems and keeps evolving to LTE-A. In the standard, an Evolved Node B (eNB) is used to represent a BS while User Equipment (UE) denotes a mobile device. An LTE network is designed to have a flat architecture without centralised network controllers. In such an architecture, eNBs are connected to core network components, the Mobility Management Entity (MME) and the Serving Gateway (S-GW), via the S1 interface. They are also interconnected through the X2 air interface, through which essential information is exchanged for distributed management. For instance, indicators are sent via the X2 air interface in Inter-cell Interference Coordination (ICIC) for Radio Resource Management (RRM). An example of the LTE architecture is given in Fig. 2.2 (a). The main reason for decentralising the intelligence is that such a

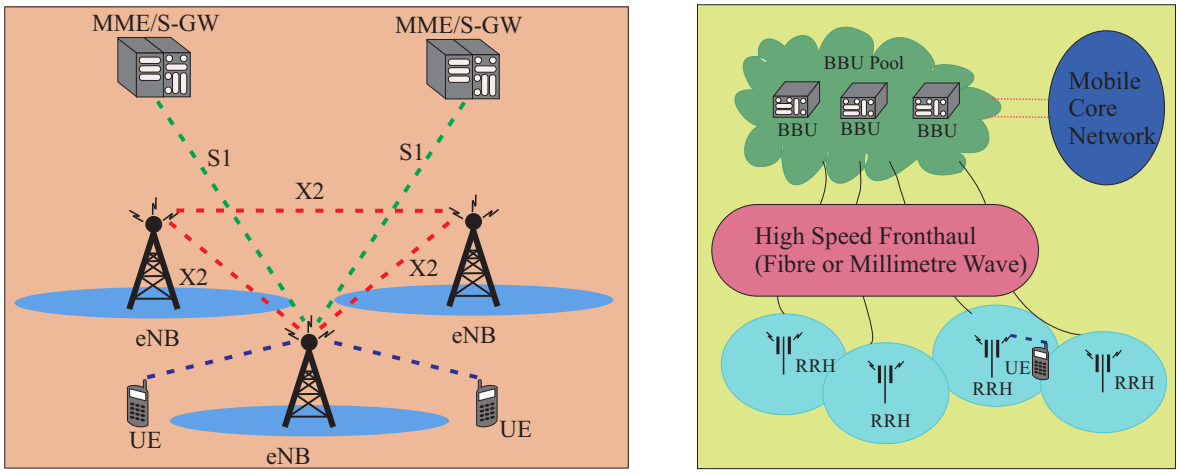
simpler network architecture can speed up connection set-up and reduce the handover time needed [14]. The LTE architecture is one of the investigated architectures in the thesis, where distributed management schemes are required.

Centralised Small Cell Networks

A large infrastructure investment is necessary in the dense deployment of small cell BSs, which bring extra capital expenditure (CAPEX). Meanwhile, the growing overall network power consumption results in an increasing OPEX and more carbon dioxide emissions. Interference management also becomes a critical issue caused by the dense deployment of small cells. In distributed architectures, collaborative RRM techniques are needed to mitigate the interference. However, it is hard to achieve the maximum performance gain with the traditional X2 interface in the current LTE architecture [15]. For the fifth generation (5G) wireless networks, some centralised architectures are re-considered due to the vital advantages compared with the distributed architecture adopted by LTE-A.

In a traditional wireless network with a centralised architecture, a centralised network controller is responsible for governing the operation, cooperation and coordination of a number of BSs within a certain area. Centralised management techniques are required for the global optimisation, such as RRM and mobility management (MM) algorithms. The thesis investigates a recently proposed disruptive centralised architecture, Cloud Radio Access Network (C-RAN, i.e. Cloud-RAN), which has attracted great attention to address the aforementioned problems. C-RAN was first proposed by the China Mobile Research Institute in 2009 and the name implies the objectives of developing centralised processing, cooperative radio, cloud, and clean (green) infrastructure RANs [16]. Since its first introduction, subsequent trials have been conducted by joint operators and academic consortia with the expectation of increasing resource utilisation, lowering energy consumption and decreasing interference [15].

Unlike BSs in the distributed architecture, remote radio heads (RRHs) in a C-RAN, also referred to remote radio units, only perform radio functions and provide signal coverage, including frequency conversion, amplification, filtering and analogue-to-digital or digital-to-analogue conversion. Despite other central controllers, a BBU pool is placed at a centralised site, where BBUs are located in close proximity as virtual BSs to better utilise processing resources and achieve the centralisation gain. For long term considerations, the BBU pool can be implemented on a cloud computing platform based on general-purpose processors with multicore and multithread techniques, which can run software to virtualise BS functionality



(a) (b)

Figure 2.2: SCN architecture overview. (a) An overview of the LTE architecture [2]. (b) An overview of the C-RAN architecture [3].

and optimise the network resource allocation [3, 17, 18]. The actual functions implemented vary. For the easiest realisation, a RRH can be simply connected to an exclusive BBU in the pool, however, missing the advantages of joint signal processing and central control. Alternatively, a set of RRHs can be linked to a central switching device, which can schedule pooled processing resources to a subset of RRHs flexibly based on demand. This approach prevails also because of a better energy efficiency and the feasibility of carrying out optimised RRM algorithms [3]. Its advantage of energy efficiency will also be discussed in later sections. A transport network, also called the fronthaul network, connects between a BBU pool and typically several dozens or several hundred of RRHs [15]. It is formed by optical fibres and/or microwave/millimetre wave links, where digitalised signals are transmitted. Although optical fibre is considered to be the most ideal backhaul for C-RAN due to the high data rate it supports, it is costly and sometimes impossible to deploy at some places. Wireless solutions, especially millimetre wave, is preferable in this situation if meeting the prerequisite that a relatively large bandwidth and a high data rate can be provided. Fig. 2.2 (b) gives an overview of the C-RAN architecture.

Thanks to the aggregation of BBUs, RRHs only keep simple functions with low cost, which reduces CAPEX. The centralised signal processing units in one site lower management cost, which lowers OPEX. By the cloudification of computation resources and the dynamic resource assignment on a pool level, the resource utilisation and power consumption can both be improved. Moreover, with software based platforms, the operation and the functionality of BBU pools can be adjusted in response to real-time dynamic radio environment based

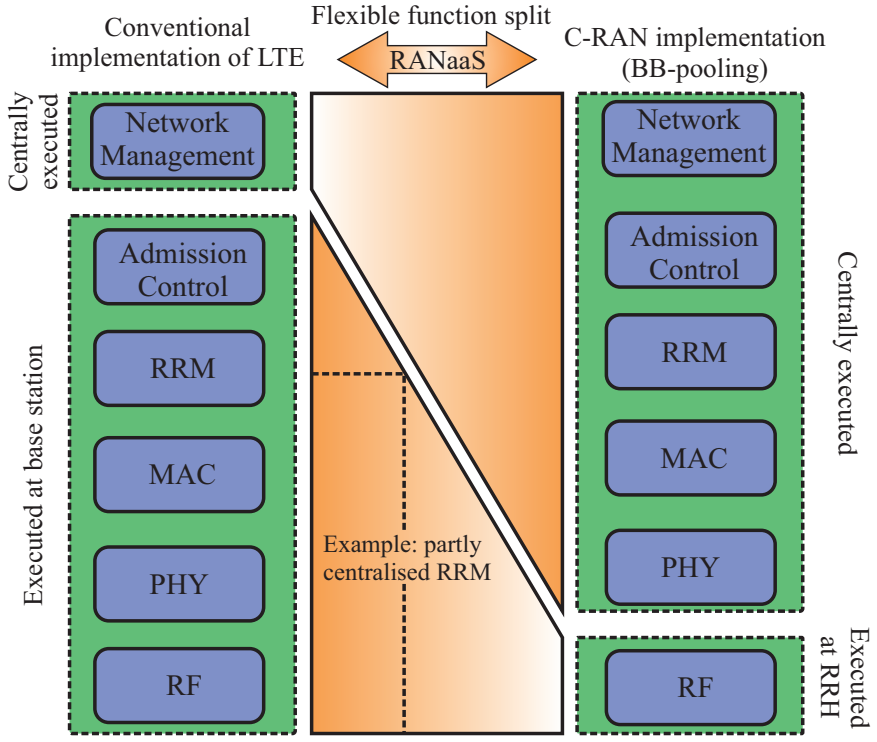


Figure 2.3: RANaaS concept with a flexible function split (directly reproduced from [4]).

on the information collected by each RRH [15, 18]. This innovates a concept, RAN as a Service (RANaaS) [19]. RANaaS flexibly splits BS functions, such as physical (PHY) layer functions, Media Access Control (MAC), RRM, admission control and network management, into a centralised platform based on cloud infrastructure. This makes the functionality of a RAN partially centralised based on actual service requirements, as shown in Fig. 2.3. C-RAN also gives birth to novel business models and services exploiting the virtues of cloudification and software based platforms [18].

On the way to the aforementioned benefits, the C-RAN architecture is still facing foreseeable challenges. The key challenge is to provide high speed and economical fronthauls for centralisation knowing that developing fibre connections is costly and wireless links have a limitations in data rate due to the restricted bandwidth. The most viable option is enabling the millimetre wave technology in fronthaul networks to achieve a high data rate when solving the problem of the fibre deployment. Another challenge is implementing virtualised network functions and network controlling algorithms over a centralised platform. This involves the study of the relative technologies such as Network Function Virtualisation (NFV) and Software Defined Networking (SDN). Management schemes different from the current distributed ones are also required for centralisation, which is considered in the thesis.

2.2.2 Heterogeneity in Small Cell Networks

In the evolution of wireless networks, different access technologies have emerged to satisfy diverse demands at the time. Giving a broader definition to SCN, it may integrate Multiple Radio Access Technologies (Multi-RATs) beyond the conventional cellular ones. By thinking about the level of densification of a ultra-dense network (UDN), envisioned Inter-site Distances (ISD) between small cell access nodes range from a few meters in indoor deployments up to roughly 50 m in outdoor deployments [20]. Therefore, SCNs also allow the existence of different types of BSs, including Relay Nodes (RNs), Pico Base Stations (PBSs), Femto Base Stations/Access Points (FBSs/FAPs) and remote radio heads (RRHs). They co-exist with the MBSs and are expected to accommodate specific requirements in different scenarios. Although BSs of the same size with a single access technology are considered in the thesis, the heterogeneity in SCNs is reviewed for completeness and to better understand the SCNs studied. The review also indicates the directions, where the work presented in the thesis can be potentially transplanted.

Multiple Radio Access Technologies

With the popularity of wireless communication, there is an increasing number of mobile devices capable of connecting to the internet via different Access Points (APs) with multiple radio communication interfaces in addition to cellular networks. They not only compete with the successive generations of cellular access technologies, but can also be integrated together in future SCNs to provide ubiquitous computing and improved capacity, enhancing the overall performance.

Wireless Local Area Network (WLAN), based on Institute of Electrical and Electronics Engineers (IEEE) 802.11 family, is a popular RAT dedicated to providing high data rates with a low cost. WLAN APs are usually developed in hotspot areas, including homes, shopping malls, hotels, university campuses and airports [21], enabling users connected to the Internet within a local coverage area. Due to the extensive application, WLAN APs are able to offload the traffic from mobile networks to some extent. Another RAT, Wireless Personal Area Network (WPAN), allows devices with only WLAN protocol to access cellular networks through another multi-protocol (for cellular networks and WLANs) device. The multi-protocol device creates a short range WLAN AP, which routes the traffic coming from other devices to the cellular networks. This may give birth to new device-to-device services

for cellular subscribers [22].

Despite the potential capacity increment contributed by the Multi-RATs, there are many challenges when deploying Multi-RAT SCNs. The first problem to be addressed is the integration of different RATs with different protocols and architectures, which are developed independently by different standardisation bodies. Specific joint RRMs have to be designed to improve the overall user experience. Network selection or vertical MM strategies should also be considered to retain seamless connection to the Internet [23].

Node Types

Motivated by the capacity benefits, both academia and industry have devoted significant efforts to the deployment of SCNs of different cell sizes to meet diverse demands in the past. Together with the MBS, the main types of small cell nodes are introduced as follows.

Macro Base Stations

MBSs are high power BSs serving conventional cellular networks. MBSs cover large area macro cells ranging up to more than 30 kilometres using high transmit power varying from 5 W to 40 W [2]. Due to the high operating power and transmit power, MBSs usually require air conditioners to cool down the environments around them to keep appropriate temperatures. Considering about the large sizes, the enormous costs and high power consumption of these BSs, they are not suitable for large scale deployment. However, MBSs still have value for the co-existence with future SCNs. Due to the large coverage areas, they can be used to provide essential coverage in rural areas. Moreover, a reduced frequency of handover is required by macrocells compared with smaller cells, which has inimitable advantages in handling high mobility users.

Pico Base Stations and Micro Base Stations

PBSs and micro BSs are both simplified MBSs of reduced operation and transmission power, smaller coverage areas and smaller BS sizes. PBSs typically serve smaller areas (200 metres radius or less) with lower transmission power than micro BSs. However, the distinctions between them are sometimes not clear. They are usually carefully planned and deployed in indoor or outdoor environments by operators with open access to all mobile users. Specifically, PBSs have transmission power varying from 250 mW to approximately 2 W for outdoor deployments and typically 100 mW or less for indoor scenarios [2].

Femto Access Points

FAPs or FBSs are commonly used to describe unplanned indoor nodes deployed by consumers depending on their own demands. They cover femto cells ranging from 10 to 50 metres with 100 mW transmit power or less [2]. FAPs are connected to digital subscriber lines or cable modems as backhaul links. They can be further classified based on whether they permit access to all devices. Open access FAPs operate like PBSs except using the aforementioned backhaul links, while closed access FAPs only grant access to closed subscriber groups. A hybrid fashion is also possible that all devices can access FAPs but with the specified subscriber groups given higher priorities.

Relay Nodes

Relay nodes are similar to PBSs, but are usually deployed at cell edges or coverage holes to provide throughput or coverage consolidation. Relay nodes use the air interface spectrum for the wireless backhaul. This brings challenges to the wireless system design if the backhaul links share the same frequencies as the access links. The alternative way is to use extra dedicated frequencies for the wireless backhaul to avoid the interference. However, this requires operators to spend more on spectrum resources from regulators. Directional antennas are typically deployed for RN backhaul links to have the highest gain in the directions pointing to the donor BSs.

Remote Radio Heads

The current hardware of cellular BSs includes power amplifiers, antennas and signal processing modules. Comparatively, only antennas and power amplifiers are concentrated in RRHs, which are distributed and spatially separated from the common signal processing units [24]. Multiple RRHs are required to connect to their shared centralised baseband processing unit through a high transmission rate medium, such as optical fibres. As low cost RRHs in a wide geographical area share the same large baseband processing pool, the concept is therefore attractive to operators which need lucrative small cell deployments without large infrastructure expenses.

2.2.3 Technical Challenges of SCNs

As shown in Fig. 2.4, a future multi-RAT SCN deployed over the macrocell will involve diverse protocols and cells of various sizes, which constitute a hierarchical and complicated network. Associated with the potential benefits, there are still many technical challenges to be addressed.

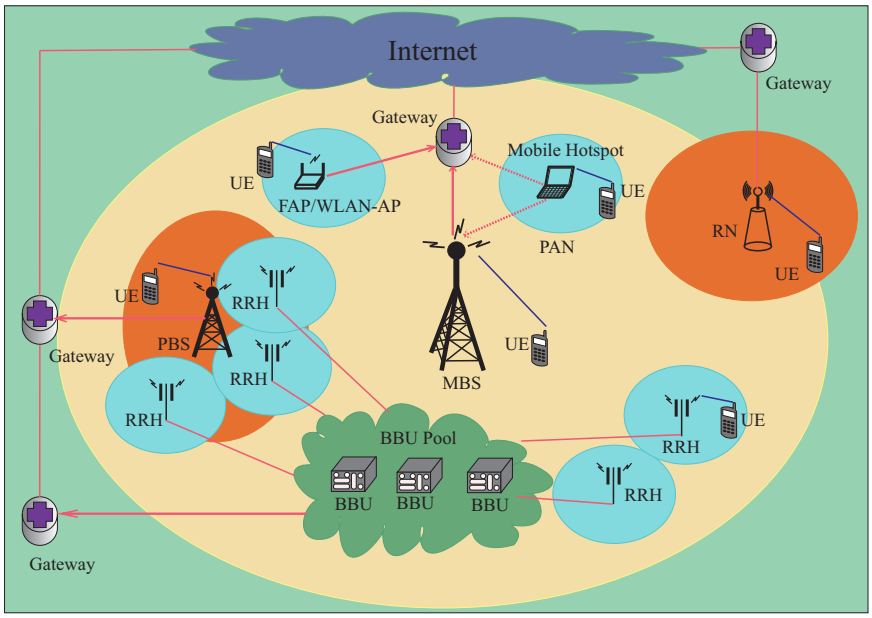


Figure 2.4: An example of heterogeneous SCN architecture.

Interference management is a key issue, which directly relates to the system performance, such as QoS. In future SCNs, co-tier intercell interference may be more severe due to densification of unplanned random small cell deployments. In addition, inter-tier intercell interference also arises as a result of overlaying co-channel SCNs with diverse levels of transmission power. Particular schemes should be considered to both vertically (across multiple tiers) and horizontally (across cells in a tier) optimise SCN operation. With the capability of supporting multi-RAT, modern mobile wireless devices may also suffer from in-device interference when transceivers in proximity are operating on the same, adjacent, harmonic/sub-harmonic frequencies [25]. Therefore, advanced interference mitigation and in-device interference cancellation techniques have to be developed.

When developing UDNs, a large scale of small cell BSs have to be deployed so that accurate cell planning is not always feasible. Moreover, with the constraints of the actual scenarios, BSs are sometimes mounted without correlating to the local service demands. Unplanned home or enterprise nodes, especially these privately owned nodes of closed access, may pose extra interference. An autonomous engine is essential to self-configure, self-optimise and self-heal the SCNs, yielding a Self-organised Network (SON). The more frequent handover in UDNs and multi-RAT scenarios have made MM critical, especially for high mobility users. Potential MM strategies, such as virtual cell, user grouping and adaptive handover may be considered for this challenge. The management of neighbour cell list (NCL) should then fit the MM strategies under the SON concept. The thesis aims at solving one of the challenges,

which is on-demand radio resource provisioning, positively correlating to the hotspot areas.

2.3 Spatial Modelling for Small Cell Networks

It is important to realistically characterise the topology of the BSs in a network since the radio power is dependent on the distance between a transmitter and a receiver, which affects the system performance. It is the prerequisite of obtaining accurate results in the performance analysis of cellular networks via a theoretical and mathematical approach due to its reliance on spatial models. It is also the primary task for simulation-based research, where layouts of the BSs should be effectively captured and will affect the subsequent performance evaluation.

The topology of the current networks has been complicated by the gradual deployment of SCNs. The future SCNs are envisioned to have more complexity and heterogeneity with multi-RATs involved. Therefore, the spatial distribution patterns of the BSs in a SCN have to be reconsidered based on their own features. This section reviews the mostly used models for cellular networks and investigates the scenarios where they are applicable. SCNs are particularly considered in the study to match the topic of the thesis.

2.3.1 Grid Based Models

In grid based models, cellular BSs are placed on a 2D grid. They have been extensively used to model the conventional macro BSs for the system-level and simulation-based research carried by both the industry and the academia [26]. In the past decades, the positions of cellular BSs were based on careful cell planning to ensure a reasonable placement. However, the actual layouts of the cellular networks at the time do not strictly follow a grid pattern even with cell planning due to the restrictions in site availability. Therefore, grid based models deliver idealised network topologies and thus the upper performance bound because the performance evaluation of the networks is substantially affected by the BS topology. Therefore, they are only more valid for modelling SCNs in some special cases including the scenarios where the small cell BSs are attached to facilities exhibiting a regular pattern, such as street lamps. The most commonly used grid based models are the hexagonal lattice and the square lattice.

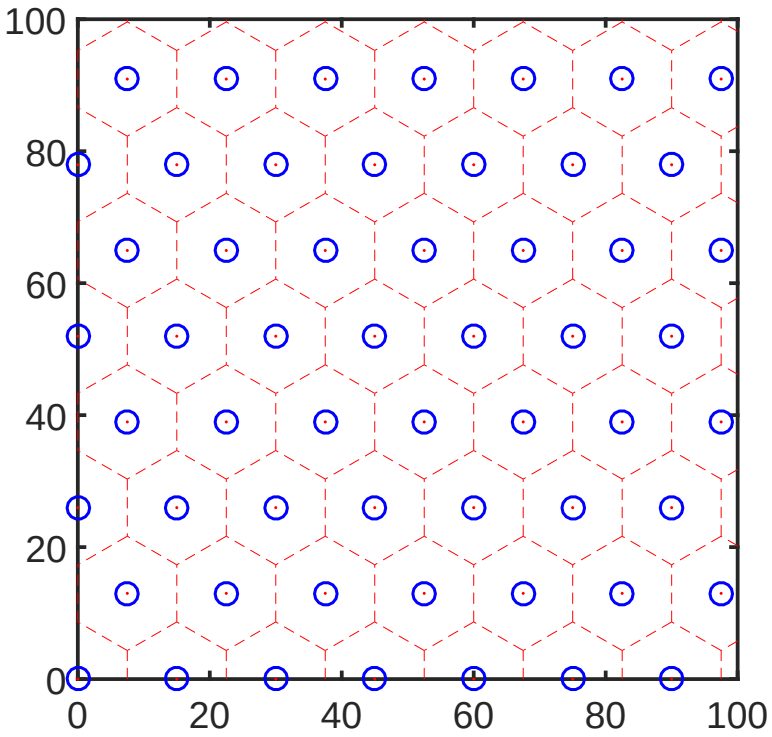


Figure 2.5: A realisation of the hexagonal lattice model. $\rho = 15$.

Hexagonal Lattice

The hexagonal lattice is one of the models used to model the spatial distribution of BSs in the thesis. Considering a 2D Euclidean plane, the hexagonal lattice $\mathcal{L}_{\mathrm{h}} \subset \mathbb{R}^2$ can be formally defined as

$$\mathcal{L}_{\mathrm{h}} = \{\rho(\frac{1}{2}k + m, \frac{\sqrt{3}}{2}k + \sqrt{3}n) : k, m, n \in \mathbb{Z}\}, \tag{2.1}$$

where the coefficient $\rho \in \mathbb{R}$ is the ISD determining the density of the hexagonal lattice. The lattice can be obtained by realising all combinations of k, m and n, which form discrete coordinates. An example is shown in Fig. 2.5, where the BSs are placed on a hexagonal lattice with $\rho = 15$. The area of each resulting Voronoi cell is $\frac{\sqrt{3}}{2}\rho^2$.

Square Lattice

The square lattice $\mathcal{L}_{\mathrm{s}} \subset \mathbb{R}^2$ can be formulated as

$$\mathcal{L}_{\mathrm{s}} = \{\rho(m, n) : m, n \in \mathbb{Z}\}, \tag{2.2}$$

where the coefficient $\rho \in \mathbb{R}$ determines the density of the square lattice in the same way and the area of a Voronoi cell, which is ρ^2. The example shown in Fig. 2.6 displays the locations

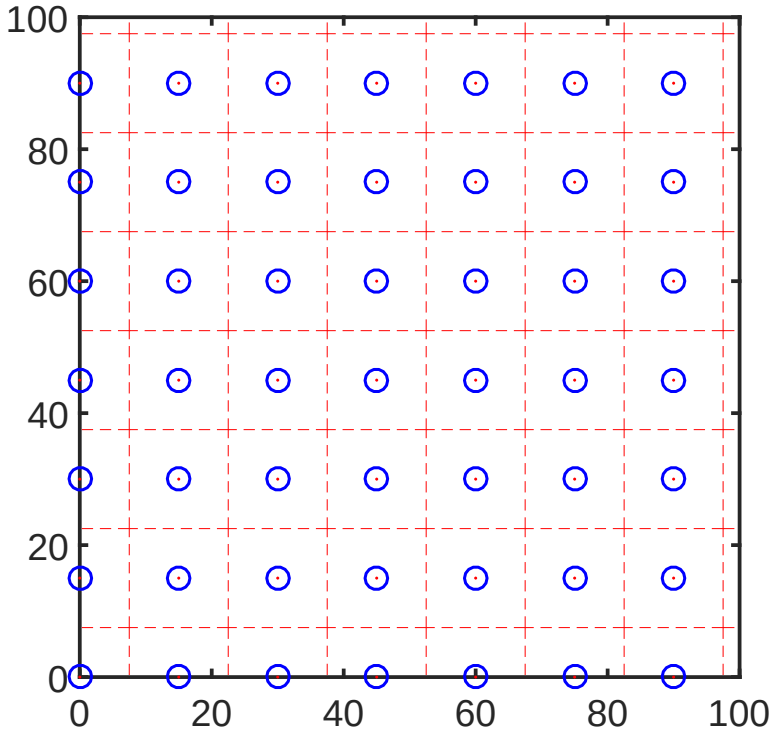


Figure 2.6: A realisation of the square lattice model. $\rho = 15$.

of BSs placed on a square lattice with $\rho = 15$.

2.3.2 Stochastic Models

With the proliferation of SCNs, more complexity is involved in modelling cellular networks. Considering the large amount of the small cell BSs, elaborated cell planning becomes less likely to be undertaken before deployment, which further limits the application of grid based models. Due to the reduced cell sizes, the deviation of the model used from the physical BS distribution has more impact on the analysis of the system performance. With grid based models, most performance metrics cannot be identified using an analytical approach [26]. Although grid based models can still be adopted for carefully planned networks, stochastic models are required to model SCNs in more general scenarios. The on-demand radio resources provisioning considered in the thesis is in order to address the power consumption problem for general SCNs, motivating the study of stochastic spacial models.

Recently, a new modelling technique based on stochastic geometry has been adopted to characterise the randomness in the geometry configuration of SCNs, i.e. the topology of small cell BSs. Stochastic geometry is a mathematical and statistical tool for modelling wireless networks, from which tractable analytical results and insights can be obtained [27,28].

Through the theoretical analysis using stochastic geometry, metrics of interest can be regarded as the average over a large number of nodes at different locations or over many network realizations. Alternatively seen from a generic node, the metrics can also be treated as the expectations weighted by probabilities [27]. For simulation-based work, stochastic models also prevail over grid based models because they can represent more general SCNs due to the randomness captured.

Stochastic models for SCNs are usually realised by spatial point processes. Considering a 2D Euclidean plane $\mathbb{R}$, informally, a spatial point process can be represented as a countable random set $\Phi = \{x_i\}$ [28], where $i \in \mathbb{N}^*$ and $\Phi \subset \mathbb{R}^2$. It is usually characterised by a random counting measure $\Phi(B)$, which denotes the number of points in a set $B \subset \mathbb{R}^2$. Note that there are many kinds of spatial point processes, which can be produced from the fundamental point processes by thinning, clustering, superposition and transforming their distributions with the application of probability densities [29]. This review only presents the ones that are most commonly utilised in the field of wireless communication for node spatial modelling, including the one extensively studied in the thesis.

Poisson Point Process

A point process $\Phi \subset \mathbb{R}^2$ is a Poisson Point Process (PPP) Φ_p if and only if $\Phi(B)$ is a Poisson random variable for every bounded closed set B and $\Phi(B_1)$, $\Phi(B_2)$, $\ldots$, $\Phi(B_m)$ are independent for disjoint B_1, B_2, $\ldots$, B_m ($m \in \mathbb{N}^*$) [30]. In other words, for a homogeneous PPP (HPPP) Φ_p with intensity λ_p, $\Phi_\mathrm{p}(B)$ is the Poisson random number with $\mathbf{E}\big(\Phi_\mathrm{p}(B)\big)$ equal to $\lambda_\mathrm{p}|B|$ for any disjoint B, where intensity λ_p describes the number of points generated in the point process per unit area and $|\cdot|$ is the Lebesgue measure. For the general PPP, $\mathbf{E}\big(\Phi_\mathrm{p}(B)\big)$ is $\int_B \lambda_\mathrm{p}(x)\mathrm{d}x$ instead if λ_p is the location-dependent intensity function. Using the general case of Φ_p, the probability of having k points in B [30] can be written as

$$\mathbf{P}\big(\Phi_\mathrm{p}(B) = k\big) = \exp\left(-\int_B \lambda_\mathrm{p}(x)\mathrm{d}x\right)\frac{\left(\int_B \lambda_\mathrm{p}(x)\mathrm{d}x\right)^k}{k!}. \tag{2.3}$$

Fig. 2.7 displays a realisation of a HPPP with $\lambda_\mathrm{p} = 0.005$.

As the most fundamental point process applied to model the distribution of the nodes in a wireless network, HPPP gains the wide popularity and has been used for large-scale ad hoc networks for more than three decades [27]. Thanks to its independence property, it

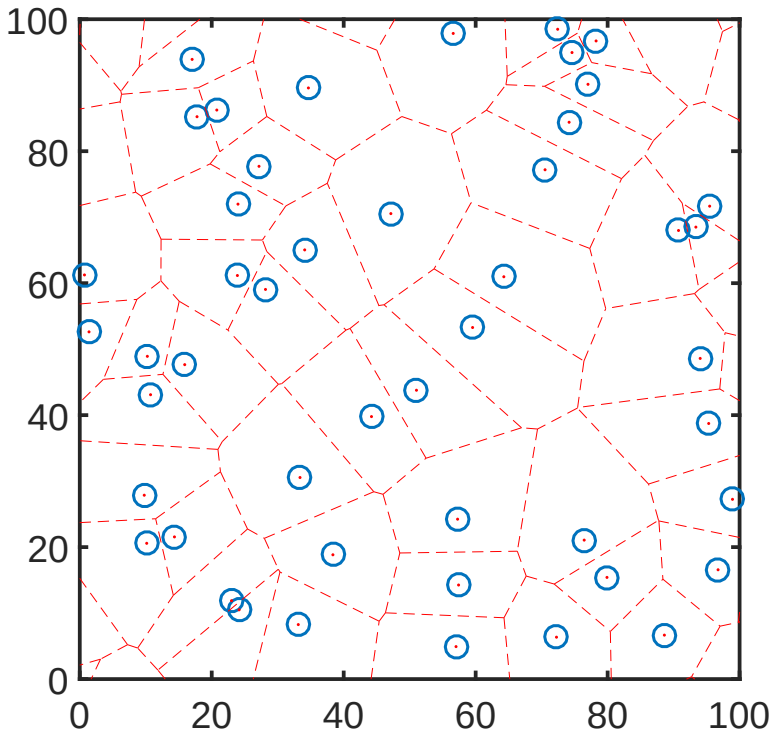


Figure 2.7: A realisation of the HPPP model. $\lambda_{\mathrm{p}} = 0.005$.

is mathematically tractable, which allows the performance of HPPP-based networks to be well understood. As it involves randomness for the node placement, it can also be used to abstract the topology of SCNs and provides tight bounds for the system performance. Its disadvantage is obvious that the BSs modelled are likely to be arbitrarily close to each other, which is not realistic.

Binomial Point Process

A Binomial Point Process (BPP) Φ_{b} is a result of identically and independently distributing fixed and finite $n \in \mathbb{N}$ points on a bounded closed set $W \subset \mathbb{R}^2$ [31]. The name of Φ_{b} comes from the fact that $\Phi_{\mathrm{b}}(B)$ for $B \subset W$ is binomially distributed with parameters $n = \Phi_{\mathrm{b}}(W)$ and p [30]. In other words,

$$\mathbf{P}\big(\Phi_{\mathrm{b}}(B) = k\big) = \binom{n}{k} p^k (1-p)^{n-k}, \tag{2.4}$$

for $k \in \mathbb{N}$, $k \leq n$, where the binomial coefficient is

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}. \tag{2.5}$$

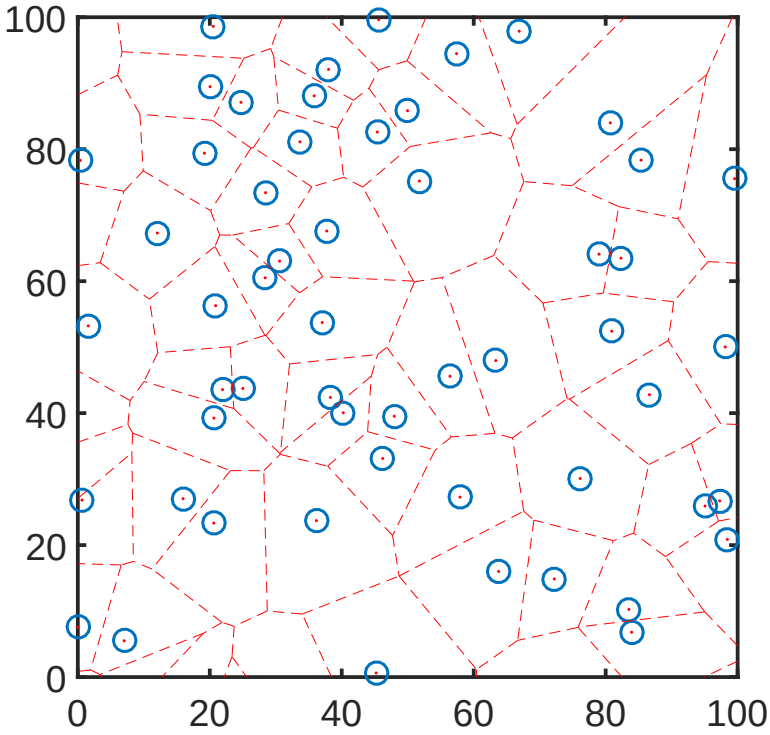


Figure 2.8: A realisation of the BPP model. $n = 50$, W is the area with the side length equal to 100.

As BPP usually indicates a uniform distribution of the points [29], $p = |B|/|W|$. In this case, $\mathbf{E}\big(\Phi_{\mathrm{b}}(B)\big) = n|B|/|W|$.

Different from a PPP, the total number of points is a fixed and finite number. It can be viewed as a conditional HPPP on the condition that the number of points is known and is n [29]. Conversely, a HPPP can be obtained by taking the number of points n to be a Poisson variable with mean n. This property yields a similar distribution pattern for the realisation of a BPP on W with $n = 50$ and $|W| = 10000$ as shown in Fig. 2.8, which cannot be visually distinguished from the the HPPP with the same intensity ($n/|W| = 0.005$). However, the fixed total number of points makes the BPP more appropriate to abstract a network in a particular area if the number of its nodes in known [31].

Poisson Hard-core Point Process

A Poisson Hard-core Point Process is a point process built on a PPP with a minimum pairwise distance δ, i.e. the distance between any two points is larger than the hard-core distance δ. These point processes are usually generated by starting with a parent PPP and then removing the unqualified points, violating the hard-core condition [30], which is called

a thinning process. It introduces a repulsion effect among the primary points generated by the parent PPP. The other similar category is soft-core process like a Strauss Process for example, where the repulsion is not 'hardly' restricted to a constant minimum distance, but relies on a probability function making points less likely to have a pairwise distance less than an interaction distance [28, 32]. As a representative, the Matérn Hard-core Point Process (MHPP) is introduced here.

The origin MHPP comprises the type I and the type II, each of which is rendered by a dependent thinning process based on a parent HPPP with intensity λ_{p}. In the thinning process of the type I, any pair of the points generated in the parent HPPP is deleted if their pairwise distance is not larger than a hard-core distance δ. While in the type II MHPP, each of the primary points is associated with an independently generated random mark that is uniformly distributed over the range (0,1). A primary point is deleted in a thinning process only if it has another primary point with a smaller mark within a distance of δ. All the primary points are simultaneously removed in the thinning process of both types meaning that the primary points to be deleted can still interact with other near points. It becomes a type III MHPP if the point removal is carried out based on a one-by-one basis. However, it is significantly less tractable than the first two types [30] and is not widely used in BS spatial modelling.

Both the type I and the type II have been extensively used to model the contention-based medium access control, like the carrier sensing multiple access scheme, which prevents close nodes from transmitting at the same time [33]. In this case, the hard-core distance is converted to the maximum interference that can be perceived. A node within the hard-core distance of another active node, i.e. having interference larger than an equivalent threshold, should stay mute. Additionally, type II MHPP has also gained increasing interest in modelling cellular BS distributions in the recent research due to its repulsion feature [27, 28, 34]. SCNs usually involve more randomness in the BS placement compared with conventional macrocell networks while they are less likely to have the BSs placed arbitrarily close to each other compared with an ad hoc network. This feature is well reflected in a MHPP, which is based on a HPPP with randomness and a hard-core distance leading to repulsion. The hard-core distance in this case considers the cell planning more or less existing in the development of SCNs.

The type II MHPP is adopted in the thesis to capture the randomness in the BS placement in SCNs due to its popularity. Therefore, Fig. 2.9 provides an example of the type II MHPP

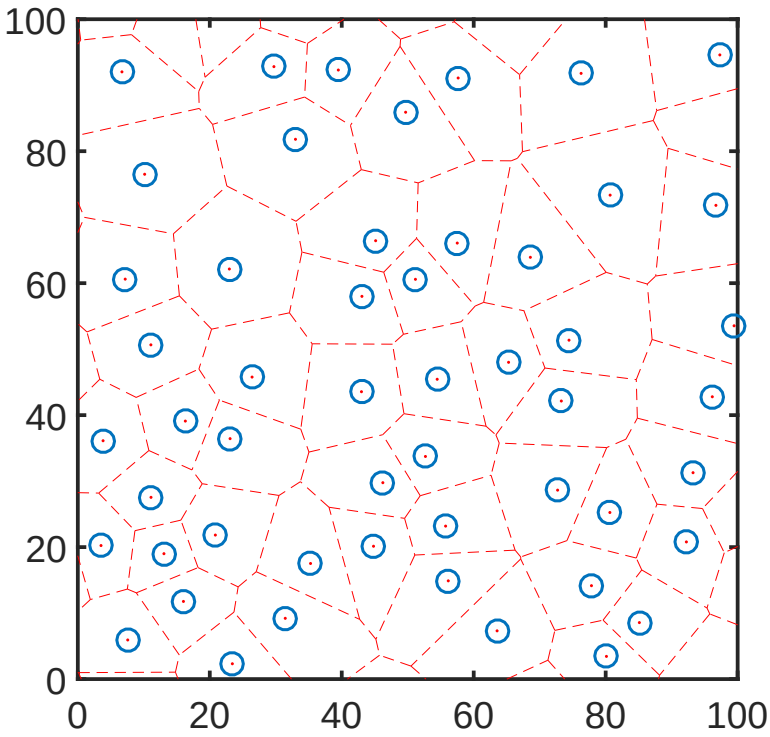


Figure 2.9: A realisation of the MHPP model. $\lambda_p = 0.01$, $\delta = 7.12$.

with $\lambda_p = 0.01$ and $\delta = 7.12$. The observed intensity is equivalent to that of the HPPP in Fig. 2.7 which has intensity of 0.005. Compared with the HPPP, the distribution of the MHPP manifests a more uniform pattern, where it is not possible to place points arbitrarily close to each other due to the hard-core distance.

Poisson Cluster Point Process

A general cluster point process is produced by first generating primary points via a parent point process and some secondary point processes, one per primary point. Then, each secondary process is translated to the positions of the primary points [30]. The final cluster point process is the union of all the secondary points. A Poisson Cluster Point Process, as the name indicates, is based on a parent HPPP with each primary point replaced by a cluster of points independently and identically distributed in the spatial domain [27, 29]. One of its special categories is defined by specifying the translation process, where the secondary points in each cluster are random in number and are scattered independently (both intra-cluster and inter-cluster) with an identical spatial probability density around its corresponding primary point. This is called Neyman-Scott Point Process [29, 30, 35]. It is worth mentioning that the centre of each cluster is shifted to a primary point but the primary points are not observed

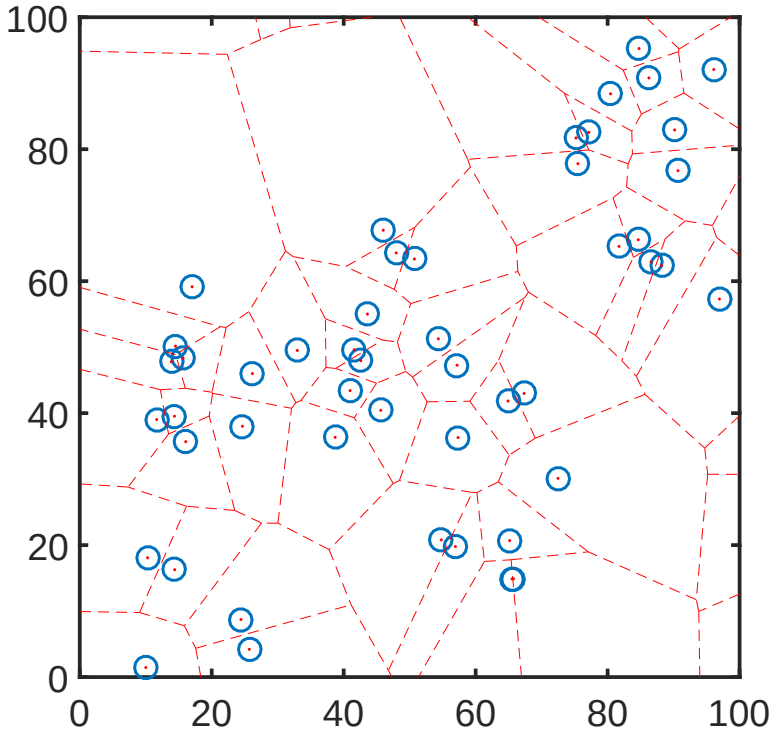


Figure 2.10: A realisation of the MCPP model. $\lambda_p = 0.001$, $\bar{c} = 5$, $R = 20$.

in the final point pattern.

A representative of the Neyman-Scott Point Process is the Matérn Cluster Point Process (MCPP). Denoting the set of primary points of a parent HPPP as $\Phi_p = \{x_i\}$ ($i \in \mathbb{N}^*$, $i \leq n$, n is the number of the primary points) with intensity λ_p, the secondary points in each cluster Φ_i^{mc} are uniformly scattered on a disc of radius R centred at x_i, where a 2D case is considered. The number of the secondary points in each cluster conforms to a Poisson distribution with mean $\bar{c}$. The final MCPP Φ_{mc} is $\bigcup_i^n \Phi_i^{mc}$, which has an overall intensity $\lambda_p \bar{c}$ [29,30,32]. There are also other Neyman-Scott Point Processes with different intensity functions for a cluster, such as a Thomas Cluster Point Process [30].

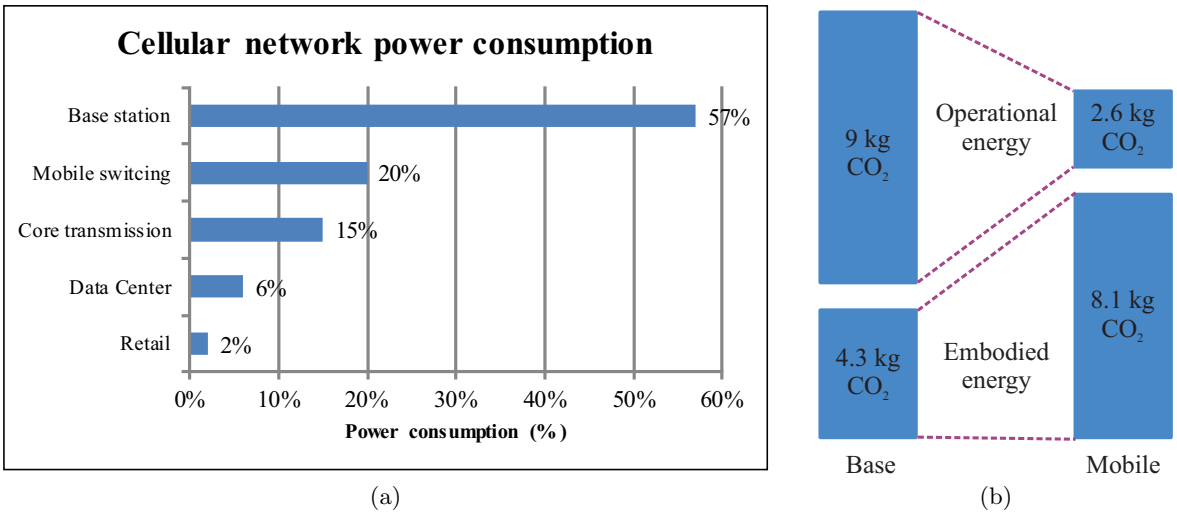
A cluster point process fundamentally reverses the repulsive thinning process, i.e. attracts points with a certain pattern. It can reflect the fact when the nodes in a SCN are clustered according to a particular social behaviour [27], e.g. the BSs are deployed in some discrete hotspot areas, but there is randomness in the deployment of BSs in each cluster. As can be seen from Fig. 2.10, the MCPP with $\lambda_p = 0.001$, $\bar{c} = 5$ and $R = 20$ has an equal intensity as the HPPP in Fig. 2.7 ($\lambda_p \bar{c} = 0.005$). The final observed points deliver an aggregation pattern at some places which may refer to the hotspot areas concerned in the planning of the clusters of BSs.

2.4 Green Networking

The advent and popularity of smart mobile devices, and the development of Information and Communication Technology (ICT) in the last decade have changed the living style of people, enabling them to access Internet services through various wireless communication systems in a more convenient way. The associated hidden effect of the evolution is the growth of the overall power consumption. It results in the increasing electricity bill and the generation of more greenhouse gases, which are directly tied to climate changes as well as other environment problems and bring about further economic or financial damage [36]. It is estimated that the ICT industry generates 2-4% of the man-made Carbon footprint [37] and mobile operators have already been among the top energy consumers [38]. The growing data demand from mobile consumers and the corresponding development of SCNs will keep bringing pressure to the wireless operators regarding power consumption and CO_2 emissions.

As the electricity bill accounts for a significant proportion of the OPEX (e.g. 18% for an European operator, 32% for an Indian operator [39]), mobile operators have to face a potential OPEX increase contributed by the large-scale deployment of SCNs. Driven by both the social responsibility and the financial purpose, green networking becomes attractive to mobile operators. However in the conventional design of communications systems, protocols and networks, academia and industry mainly focus on providing ubiquitous access and large capacity while the power consumption aspect has not been greatly considered [37]. Therefore, it is necessary to emphasise green networking when designing the next generation wireless systems, not only to reduce the overall power consumption, but also to minimise the greenhouse gas emissions.

On-demand radio resource provisioning studied in the thesis is a subset of green networking technologies, which are introduced in this section. The first step of saving energy in a wireless network is to understand what elements of a RAN consume the most energy and where the largest improvements can be achieved, after which suitable network architectures can be identified and the effort can be devoted at the right places. In Fig. 2.11 (a), the power consumption of operating a typical cellular wireless network is presented, where more than 50% of the total power is consumed by BSs. In Fig. 2.11 (b), it is shown that the operational energy is the main contributor at the BS side while the embodied energy (consumed during the manufacturing process) of mobile devices is more significant. The main reason for this difference is the limited lifetime of mobile devices (typically 2 years compared with 10-15



(a) (b)

Figure 2.11: Network power consumption and environmental impact. (a) Typical wireless network power consumption (reproduced from [5]). (b) CO_2 emissions per subscriber per year (reproduced from [6]).

years for BSs) and the small recycling proportion of mobile devices [40]. Both figures indicate that there is a large energy saving potential lying in the operation of BSs. This section reviews some important technologies for the reduction in operating power consumption of BSs and introduces the state-of-art of on-demand radio resource provisioning.

2.4.1 Enabling Technologies for Green Networking

As an integrated and comprehensive research area, green networking covers all layers of the protocol stack or the network architectures. The target of green networking is to balance the fundamental trade-offs between the energy saving aspect and the overall system performance [41]. To address the challenge, it is essential to consider various techniques such as energy-efficient BS deployment, energy-efficient protocol and architecture design, opportunistic network access, energy-efficient RRM, cross-layer design, cognitive and cooperative network deployment, HetSNet deployment, smart grids [38,42]. In the following context, the green technologies related to the scope of the thesis are introduced.

Energy Saving Potential in BSs

To meet the rapidly growing data traffic, the number of BSs has increased tremendously in the past years and will keep growing continuously. As aforementioned, the power consumption of BSs is the main contributor to the total operating power consumption of a network. This

accounts for the significant increase of energy cost or OPEX of mobile operators so that it is crucial to improve the energy efficiency of BSs. This aspect relates to the parameters of the power consumption models in Section 3.2.7. The main approaches for improvements include the minimisation of the hardware energy consumption of BSs and the application of advanced power management algorithms, such as sleep mode control algorithms. Renewable energy resources can also be adopted in some scenarios if they are financially worth investing.

One of the energy saving approaches at the BS side is to improve the energy efficiency of BS hardware modules, e.g. to make power consumption more linear to the BS load. Fig. 2.12 presents a typical structure of a BS site. A BS contains power amplifiers (PAs), RF modules, feeders and the correspondingly served antenna interfaces (AIs), which are all connected to a BBU. They are powered by a DC-DC power supply and the whole site is powered by a main supply (AC-DC unit). There is a cooling system sometimes depending on the BS type for the purpose of climate control. Among all the hardware modules, PAs account for most of the total power consumption, which is dependent on the frequency band, the used modulation scheme and the operating environment [43]. Most of the power consumed in the currently adopted PAs is wasted as heat, which requires an additional cooling system and therefore costs more energy. The efficiency of PAs ranges from 5% to 20%. It is expected to reach around 70% based on a switch-mode PA architecture and more efficient modulation schemes [44], which are promising approaches to enhance the overall hardware energy efficiency. It is worth mentioning that there is no cooling system for small-size and low-power BSs (such as FBSs). They are cooled down by the natural environment, which is an advantage of small cell BSs in terms of power consumption.

Another approach is to use network level power management schemes. Specifically, sleep modes can be designed to exploit the dynamic network traffic loads varying in space and time. By adopting transceiver sleep modes, one or multiple transceivers of a BS can be switched off when it is not necessary to have so many transceivers serving a relatively low-level local traffic load. If the traffic load of a cell is even lower, it is possible to switch the whole BS to a sleep mode without its radio function activated and its coverage area can be compensated for by its neighbour BSs. The power management schemes can either be realised in a centralised or a distributed cooperative way. Other power management schemes that control component power consumption at various load levels can also be implemented. More details about on-demand radio resource provisioning and sleep mode operation are presented in the later context.

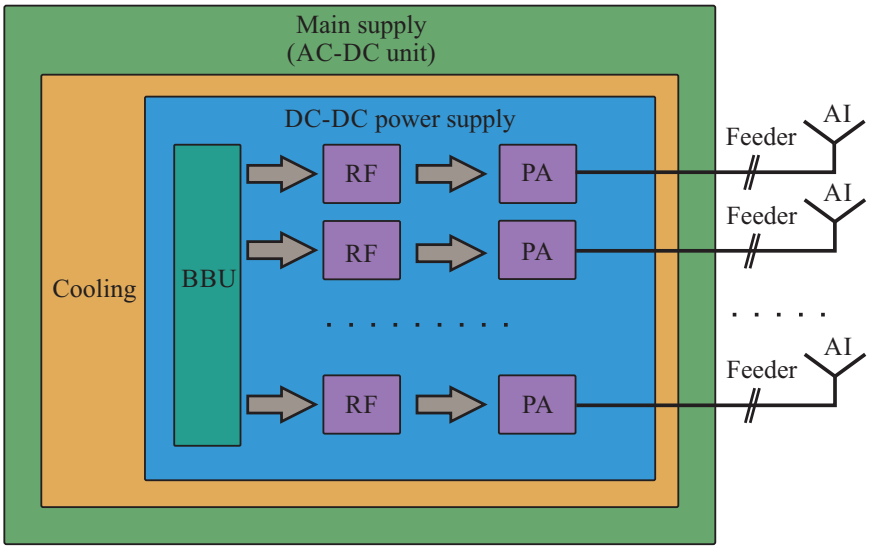


Figure 2.12: Hardware modules of a typical base station site [7].

Energy-efficient Network Architecuture

As introduced before, the deployment of HetSNets is an effective way to deliver ultra-high capacity density wireless services. From the perspective of power consumption, HetSNets also contribute to energy saving at a network level. In network densification, the deployment of high power and big size MBSs is both cost-inefficient and energy-inefficient. Instead, small cell BSs, as low power entities, can reduce the average operating power consumption due to their sizes while providing the same amount of radio resources. For example, a typical FBS only draws 5 W compared with 5 kW needed for a MBS [42]. However, it is over-qualified to deploy additional small cell BSs even at areas with low traffic demands, e.g. rural areas. In this scenario, MBSs or medium power BSs can be chosen to cover wide areas and ensure the availability of wireless services. Hence, the deployment of HetSNets with mixed BSs becomes a promising solution to boosting capacity while achieving energy efficiency [45, 46]. As a result, it adds more complexity into cell planning, where the strategy of choosing appropriate types of BSs for different locations becomes a vital part.

Considering the SCN architecture of a single layer of a HetSNet, C-RAN is given much expectation for the reduction in network power consumption. As a kind of small cell BS, RRHs also feature the low operating power and low transmit power. Moreover, they can also be fully switched off when there is a low local traffic level by more efficient and optimised centralised schemes. Efficient cooling systems can be assembled at centralised BBU pools to save energy [47]. The BBU pools can also be implemented with a management scheme of the baseband processing capability, where the total number of active processing units can be

reduced for energy saving when the network load is low [48]. The architecture of C-RAN is considered with schemes proposed in Chapter 7.

2.4.2 On-demand Radio Resource Provisioning

A network traffic measurement of an Italian operator in [49] shows that the service demands of users vary through a day or a week and also based on locations. The network traffic loads therefore distinguish among different periods and locations with peak and off-peak values and there is a long time when the traffic loads are at low levels. This indicates that different patterns of user activities engender spatial and temporal fluctuations in traffic distribution. In the future SCN scenario, there will be greater variation among different cells due to smaller cell sizes. In current wireless networks, the traffic load variations are not greatly considered for energy and cost saving where all network elements typically stay operating with all functions active and the whole network consumes similar power to that at a full load. To significantly reduce power consumption based on this feature of the traffic distribution, efforts have been made to adjust the network power consumption more proportional to the overall traffic load, i.e. to provision just enough radio resources on a demand basis. This field has gone in two different directions, which are manufacturing devices with more load proportional power consumption and making an entire network consume less power when the network load is at a low level [48]. The thesis focuses on the latter approach as the hypothesis indicates, i.e. on-demand radio resource provisioning involving operation of sleep modes.

Sleep Modes

Sleep modes are low power BS states where part or all of the hardware modules of a radio transceiver or a BS are shut down. Contributed by the deactivated components during certain periods, the average power consumption of a BS is reduced. The actual amount of energy saved depends on how deep the sleep mode is, where deeper sleep modes signify less functionality as well as lower power consumption [48]. Transition from a sleep mode to the normal full functionality of a transceiver or a BS requires a wake-up time, which is longer for a deeper sleep mode. When implementing sleep modes, operators should be aware of the impact of wake-up times and the corresponding energy consumed during the state transitions, which may affect the way sleep mode operation schemes are realised. A longer wake-up time or a large amount of wake-up energy consumption may limit the switching frequency of a BS so that it can not react timely to fast traffic variations or burst events, which are common in

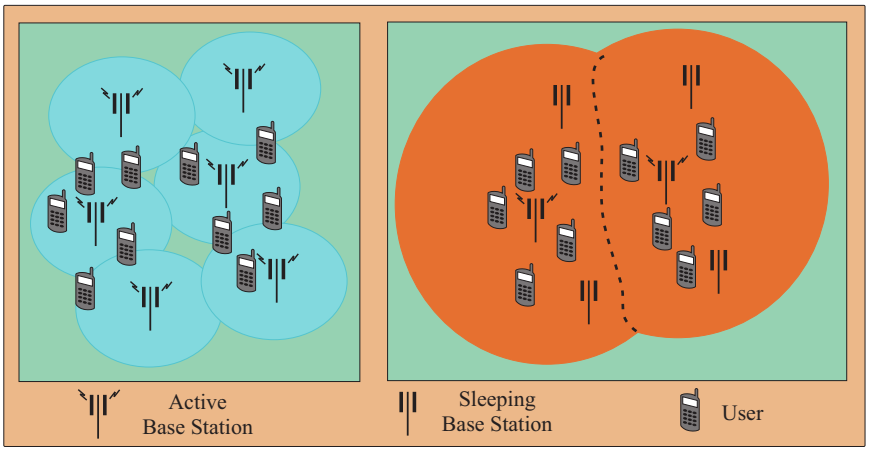


Figure 2.13: On-demand radio resource provisioning by employing sleep modes.

SCNs.

Most current BSs deployed are designed with only switch-on and switch-off states. In the future small cell BSs, the hardware design for sleep modes should accommodate the traffic properties of SCNs, which requires a short wake-up time and low wake-up energy consumption. Moreover, a BS can also be designed to have several sleep modes with different configurations meeting various network conditions. The robustness of hardware components needs to be concerned to sustain a certain frequency of state changes [48]. Extra interfaces linking the software and the controlled hardware are necessary to enable a particular sleep mode operation scheme to be easily implemented.

With properly designed sleep mode operation algorithms and BSs, a fraction of BSs in a network can be switched to the low power modes and their previous radio coverage areas are compensated by other active BSs. The thesis chooses this approach to improve on-demand radio resource provisioning. An example of this network operation is shown in Fig. 2.13. It is extremely critical that neighbour BSs of sleeping BSs can extend their coverage areas properly in order not to create a coverage hole. Capacity and QoS delivered by the whole network should be maintained when considering the maximisation of the reduction in network power consumption for dynamic traffic distributions. Additional system overheads, such as the signalling overheads and the switch on-off overheads, have to be minimised. Handover configurations of the active cells also need to be adjusted. With fewer active BSs in a network, the average radio propagation distance between BSs and users is increased, which leads to more transmission power at the user end. All of these aspects need to be taken into consideration and trade-offs have to be made to balance the pros and cons.

2.4.3 Sleep Mode Operation Algorithms

The objective of the thesis is to exploit appropriate traffic information for sleep mode operation. Thus, different sleep mode operation algorithms are reviewed and classified.

For a sleep mode operation algorithm, it can be manually configured as a static scheme (or called an offline algorithm) for a statistical traffic cycle based on pre-measured information. This type of algorithms such as [50, 51] performs well when the traffic pattern is relatively static and is accurately captured and modelled. They may be applied with different variations at different locations of a network, where the superiority of low complexity and low processing overhead should be guaranteed. Otherwise, it is better to realise the sleep mode operation in an adaptive way based on some monitored information related to the required metrics because of unusual events and unpredicted traffic variations that are especially common in SCNs. This kind of algorithms refers to online algorithms, which can be further classified into slow reaction algorithms and fast reaction algorithms depending on the frequency of making network adjustments [23, 48]. For BSs with a large wake-up time or large wake-up energy consumption, long-term measurements can be performed, with a statistical traffic perception as a result. The long-term information can be taken as the decision making material when controlling the sleep mode operation to reduce the frequency of state transition. This complies with slow reaction algorithms, e.g. [52, 53]. Conversely, a network can track the real-time traffic and react more timely to the sudden changes if the system hardware can support a swifter and more energy-efficient wake-up without letting users perceive the adjustments. In this case, fast reaction algorithms like [54, 55] can be applied. As the number of users in a conventional large macrocell is usually big, which yields statistical characteristics, the traffic volumes in different periods usually conform to a long-term distribution. For this scenario, static or slow reaction sleep mode operation schemes may be enough to respond to the traffic variations. While in SCNs, fast varying traffic volumes in each cell can be observed due to the significant reduction in the number of users in each cell, which results in a much more unbalanced traffic distribution among the small cells. This engenders the need of fast reaction algorithms to control BS state transitions, which is expected to be allowed by light-weight small cell BSs in SCNs.

In terms of the locations of decision making and control engines, sleep mode operation schemes can be categorised into centralised, distributed and partially distributed schemes. In a centralised scheme like [56,57], a central network controller is responsible to collect information measured at other network elements, i.e. BSs or users. The information is analysed and

decisions of BS state transitions are made by the central controller. The centralised architecture enables centralised algorithms, which optimise networks from a global view. However, the information collection may yield more system overheads in backhaul links. In contrast, distributed algorithms such as [54, 55] are suitable for a flat architecture, where local information from users or other BSs is collected by each BS. BSs then reason the local traffic conditions and make decisions. Although this approach loses the global view and may under-optimise the BS state transitions, it has the merits of fewer overheads and lower latency. In partially distributed schemes such as [57,58], some of the BSs in a network are given priorities to be active and control the rest of the BSs falling into their respective coverage areas. With a better view of the network conditions, the priority BSs can therefore make more reliable decisions. The first two types are usually used to describe a single-tier sleep mode operation algorithm while the partially distributed schemes can also be implemented in a HetSNet with a hierarchical architecture. The lower tier BSs with larger coverage areas can play as priority BSs, which control the state transitions of higher tier smaller BSs. Beyond these types of schemes, operators can also cooperate with each other to share the infrastructures and save energy cost.

From another point of view, sleep mode operation algorithms can also be classified depending on what kind of network elements control the operation. This classification focuses on the types of decision making engines and the functionality required by a sleep mode, which directly links to the actual power consumption of the sleep mode. The state transitions of a BS can be controlled by itself, meaning that it requires some processing hardware modules and RF modules to be active for the purposes of measuring even when they are at sleep modes, collecting and processing information [59]. These require more BS components to operate as usual or close to full power. BS state transitions can also be determined by other active BSs. This method only needs sleep modes to have a limited processing capability and lower power consumption. These kinds of sleep modes just have to be able to receive and recognise activation messages from other active BSs for wake-up [60]. Similarly to this, state transitions may be controlled by users. The difference is that the computation responsibility is distributed among users [61]. For a centralised architecture, a decision making engine can be located at a centralised site, where information is collected from the governed BSs which take the measurements. A sleep mode in this case requires fewer functions available and fewer hardware modules operating because the sleeping BSs can be controlled by the centralised controller via backhaul links. The exception to centralised architectures are the ones involving wireless

backhaul links. This case requires some hardware modules to be active similarly as the case where the state transitions of a BS are controlled by other network elements. For different BS types, appropriate strategies can be chosen. In a more comprehensive network architecture with different network conditions (e.g. various BS types, centralisation/decentralisation), mixed strategies can be taken to accommodate different requirements.

In this fast developing field, some sleep mode operation algorithms have been proposed. Some example algorithms are cited when introducing the taxonomy. A more comprehensive summary about the existing on-demand sleep mode operation algorithms is included in [48].

2.5 Summary

This chapter has reviewed some background information and established studies related to the scope of the thesis. The literature review has been focused on SCNs, which are expected to boost system capacity by enhanced frequency reuse for the next generation high-capacity wireless networks. The architectures of LTE and C-RAN are introduced as two representatives for the distributed and the centralised network architectures, respectively. The SCNs of the LTE and C-RAN architectures form the networks investigated in the thesis. The constituents of SCNs including different types of BSs and Multi-RATs with their applications have also been considered for the HetSNet deployment. The technical challenges of SCNs have been presented.

The modelling techniques for SCNs have been investigated. The hexagonal lattice and the square lattice have been introduced as examples of grid based models, which are conventional, but less preferable for modelling SCNs. The randomness in small cell BS distributions has spurred the popularity of stochastic models. The HPPP model is mostly used because of its tractability. A variation with a fixed number of points is also introduced. As the fundamental point process, HPPP can be translated with a thinning or a clustering secondary process, showing a repulsion or an aggregation effect. MHPP yielded by repulsion is the most important tool used in the thesis for spatial modelling of BSs. The application scenarios and examples of all the introduced models are presented.

Including the network capacity improvement, another important aspect of the future networks is the energy efficiency. The potential reduction in power consumption has been discussed from different perspectives due to their relevance to the scope of the thesis. Energy saving in BSs and future network architectures has been investigated. The concept of

on-demand radio resource provisioning has been introduced with sleep mode applications principally discussed. As the key research topic, sleep mode operation algorithms have been classified and their features have been studied.

Chapter 3

Methodology

Contents

3.1 Introduction

On-demand radio resource provisioning involving sleep mode operation is expected as an important concept for the development of future green SCNs of different architectures. This

thesis proposes a number of sleep mode operation algorithms for the reduction in network power consumption. The purpose of this chapter is to introduce the methodology used in the study of the algorithms, which is mainly formed by simulation based on statistical models. For the simulation, system models need to be appropriately determined to capture the practical conditions of SCNs. Related performance should be evaluated after the simulation to gauge the effectiveness of a sleep mode operation algorithm.

In Section 3.2, the system models for SCNs are presented as the basis of the simulation, which is introduced in Section 3.3. The system performance metrics related to sleep mode operation are presented in Section 3.4. Finally, the chapter is summarised in Section 3.5.

3.2 System Models

System models are presented in this section to capture the characteristics of the wireless system investigated. In the system models, the spatial models are to reflect the distributions of transmitters and receivers in a SCN, which affects the overall system performance. The antenna model and the radio propagation model is related to the attenuation of radio waves. A basic model of spectrum allocation is needed, which is a vital part of RRM and directly links to the system capacity and QoS. The link model is for the calculation of transmission data rates required for the evaluation of link level performance.

3.2.1 Spatial Modelling and Neighbour Cell List

In the research presented in the thesis, SCNs are considered to have an ultra-high density for delivering ultra-high system capacity. As SCNs can be deployed and allocated with an individual spectrum band, which will not cause interference with networks of other tiers, only a single tier ultra-dense SCN is modelled and analysed. In the spatial modelling, two scenarios are considered. The ideal spatial distribution appears in the early work presented in the thesis and has relatively regular distributions of BSs and users for modelling special cases. Whereas in the scenario with non-ideal spatial distributions, the topology of BSs is not regular, which captures the randomness in more general cases of SCNs. Users are of an unbalanced distribution to reflect the existing hotspot areas in reality. The Neighbour Cell Lists (NCLs) are also defined based on the topology of BSs.

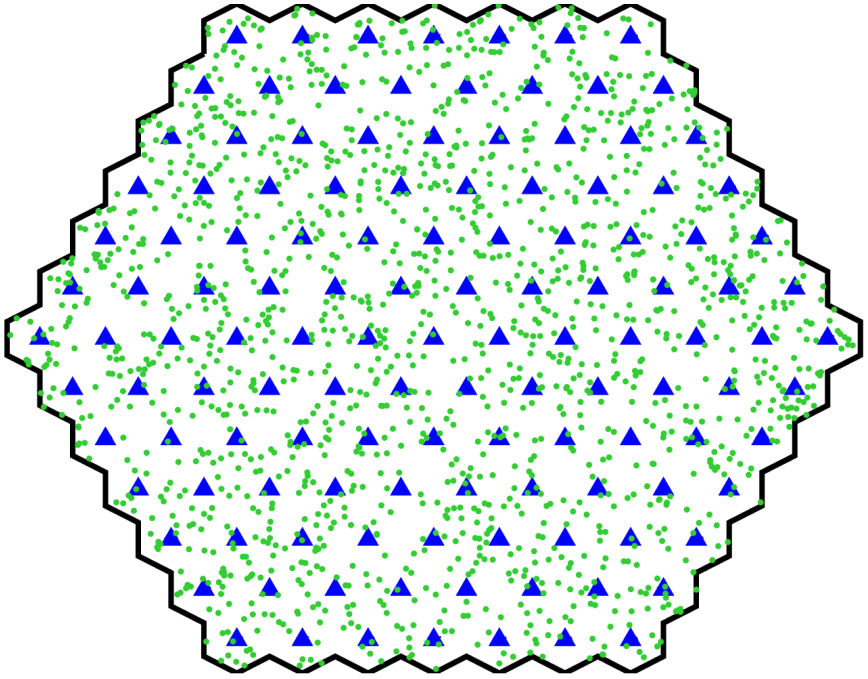


Figure 3.1: A network example with ideal spatial distributions. The BSs are depicted as the triangles. The system boundary is shown as the bold polygon. The users are depicted as the dots.

Ideal Spatial Distribution

The ideal SCN topology investigated is of a conventional hexagonal layout. This layout uses an ideal base station distribution, which usually only exists in special cases with careful cell planning. The inter-site distance (ISD) should be set to realise an ultra-dense deployment of small cells. The system is constructed within a limited area, covered by the hexagonal grids and denoted by $\mathcal{A}$, where $\mathcal{A} \subset \mathbb{R}^2$. The NCL of a target BS is defined to contain the closest six neighbour BSs around the target BS. The users conform to a uniform distribution with intensity $0.15/\mathrm{m}^2$ and are truncated into the system boundary. If an identical traffic model is applied to each of the users, the traffic in the system is relatively balanced, which is an ideal case. Fig. 3.1 shows a layout example consisting of 127 BSs (depicted as the triangles) with the system boundary shown as the bold polygon. The users are shown as the dots.

Non-ideal Spatial Distribution

The aforementioned network layout is of a grid topology. This network model is too idealised, especially for future SCNs. More random placement should be introduced to model the scenarios where there are places impossible to deploy BSs according to planned grids. To introduce the randomness in the actual BS deployment in practical networks, a homogeneous

Poisson Point Process (HPPP) is adopted as a stochastic tool to model the spatial distribution of the BSs in SCNs [62]. Following a 2D HPPP Φ_{p} with intensity λ_{p}, BSs $\mathcal{B}$ conform to a uniform distribution in the area $\mathcal{A}$, where $\mathcal{A} \subset \mathbb{R}^2$ is a bounded square area. The cardinality of $\mathcal{B}$ (the number of the BSs), $|\mathcal{B}|$, is a Poisson random variable with mean $\lambda_{\mathrm{p}}|\mathcal{A}|$, where $|\mathcal{A}|$ is the Lebesgue measure of the set $\mathcal{A}$ (the area represented by $\mathcal{A}$) [63]. However, as a significant weakness, the independence of the points in the HPPP may allow BSs to be located arbitrarily close to each other, which is not practical in reality. To solve this problem, specific repulsive dependent thinning is applied to the HPPP [34], rendering a type II Matérn Hard-core Point Process (MHPP) Φ_{m} as a result [34, 64, 65]. In the thinning process, a random mark is independently associated with each primary point of the parent HPPP and a primary point is deleted if there is another point with a smaller mark within the hard-core distance δ [65], which yields λ_{m}, intensity of Φ_{m} as

$$\lambda_{\mathrm{m}} = \frac{1 - \exp(-\lambda_{\mathrm{p}}\pi\delta^2)}{\pi\delta^2} \tag{3.1}$$

Finally, $\overline{|\mathcal{B}|}$, the cardinality mean of $\mathcal{B}$ can be calculated as:

$$\overline{|\mathcal{B}|} = \frac{|\mathcal{A}|\left[1 - \exp(-\lambda_{\mathrm{p}}\pi\delta^2)\right]}{\pi\delta^2} \tag{3.2}$$

The MHPP model will be studied with more details in Chapter 4. As the number of BSs may vary in a MHHP for different realisations, the system performance should be evaluated over a sufficient number of instances, e.g. 100 instances. In Fig. 3.2, a realisation of the SCN based on a MHPP is shown, where the small circles representing BSs are scattered in the square area $\mathcal{A}$ with the side length of 100 metres. In the MHPP, λ_{p} is set to $20000/\mathrm{km}^2$ and δ is configured to 5 metres, respectively, to suit the scenario of ultra-small cells. As can be seen, the distribution of the BSs does not follow a grid pattern, which is a non-ideal case, but closer to the reality. The other aspects of the figure are illustrated in the following paragraphs.

In recent practice, a Neighbour Cell List (NCL) of a BS is manually configured when it is deployed, which has many drawbacks due to the dynamics of the radio propagation environment [66]. Furthermore, to construct a self-organised network as one of the future requirements, a newly deployed BS has to be capable of automatically self-configuring the NCL and continuously self-optimising it. Among the most common algorithms for this purpose, a BS autonomously scans the broadcast pilot signals from adjacent BSs and creates an initial

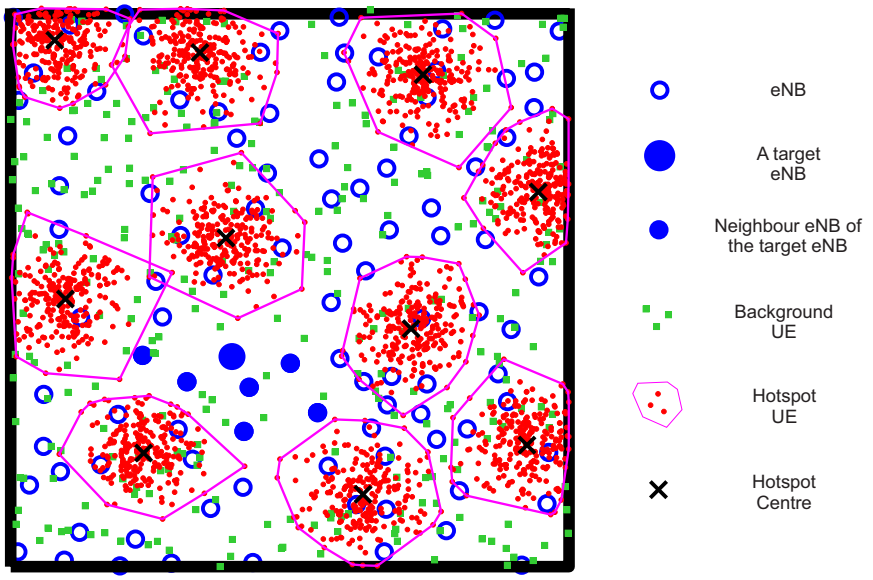


Figure 3.2: A network example with non-ideal spatial distributions. The side length of the square area $\mathcal{A}$ is 100 metres. The small circles represent the BSs. The small solid circles are the neighbour BSs of the big solid circle. Background users are depicted as squares and 10 groups of hotspot users are depicted as dots. Hotspot users in each hotspot group are encircled by the corresponding convex hull depicted by a polygon. The crosses are the centres of the hotspot groups.

NCL based on the signal-to-interference-plus-noise ratio (SINR) [67, 68], after which the BS maintains the NCL based on the SINR and other information collected by users [68, 69]. Since the self-organisation of the NCL goes beyond the scope of the study in this thesis, it is assumed that each BS $B_i \in \mathcal{B}$ ($i \in \mathbb{N}^*$) initialises the NCL $\mathcal{N}_i$ of B_i based on the pilot signals and the NCL is not updated during the lifetime of B_i. Specifically, B_i determines the six strongest pilot signals and selects the corresponding six BSs as the neighbour BSs, which is equivalent to seeking the six BSs with the lowest path loss. These six neighbour BSs constitute the NCL $\mathcal{N}_i = \{N_{i,1}, N_{i,2}, \ldots, N_{i,6}\}$, which has the same size as the one in an ideal layout based on a hexagonal grid. This strategy is used in a distributed algorithm design, where some messages need to be transmitted to neighbour BSs. In Fig. 3.2 for example, the BSs marked by the solid small circles are added to the NCL of the BS denoted by the big solid circle.

To reflect the spatially heterogeneous distribution of the users in reality, the users are modelled by two tiers, the background tier and the hotspot tier, with a hotspot ratio γ. The modelling approach is to create hotspot areas appearing in reality, but not to derive an accurate model for specific users, which falls out of the scope of the thesis. The average density of active users is determined to be $0.15/\mathrm{m}^2$. In the background tier, $1-\gamma$ of the total

active users are scattered randomly and uniformly in the system area $\mathcal{A}$ as the background users. In the hotspot tier, 2γ of the total active users are generated and further equally divided into several hotspot groups. The users in each group independently conform to a 2D normal distribution, which is truncated into $\mathcal{A}$ [70]. The probability density functions (pdfs) of the normal distributions in the XY-plane before truncation are

$$f(x, \mu_{\mathrm{X}}, \sigma_{\mathrm{n}}) = \frac{1}{\sigma_{\mathrm{n}}\sqrt{2\pi}} \exp(-\frac{(x-\mu_{\mathrm{X}})^2}{2{\sigma_{\mathrm{n}}}^2}), \tag{3.3a}$$

$$f(y, \mu_{\mathrm{Y}}, \sigma_{\mathrm{n}}) = \frac{1}{\sigma_{\mathrm{n}}\sqrt{2\pi}} \exp(-\frac{(y-\mu_{\mathrm{Y}})^2}{2{\sigma_{\mathrm{n}}}^2}). \tag{3.3b}$$

The normal distribution of each hotspot group is centred at a group centre denoted by $(\mu_{\mathrm{X}}, \mu_{\mathrm{Y}})$ with a standard deviation σ_{n} (set to 5 metres). Each group centre is randomly and uniformly scattered in $\mathcal{A}$. It keeps at least σ_{n} away from the system boundary and δ_{h} metres away from other group centres. The number of the hotspot groups are fixed to 10 and δ_{h} is set to 25 metres. To create temporal variations of the hotspot traffic, an observation of the system performance is divided into 10 periods, during each of which half of the hotspot groups (5 groups) are randomly chosen to be active with arriving traffic while the others remain inactive. In this way, the total number of active users in the SCN is fixed at any time during the observation. The purpose of the approach is to generate significant temporal variations in a simulation period to test the adaptability of an algorithm instead of having just a static snapshot of a SCN. As shown in the example in Fig. 3.2 which is produced by setting γ to 0.8, the background users are depicted as squares. The hotspot users are represented as dots with the centres of their own hotspot groups represented as the crosses. The convex hulls of the point sets of the hotspot groups are depicted as the polygons for the recognition of each hotspot group. With a spatial heterogeneity in the user distribution, areas with active hotspot users will have intensive arrived traffic, creating a non-ideal traffic distribution. It adds more complexity in the system, but can better reflect the irregular behaviours of the real users.

3.2.2 Antenna Model

Each BS is assumed to have multiple antennas like most of the modern BSs. However, a single transceiver chain with a 2D omnidirectional radiation pattern of 5 dB gain is assumed for transceiving since the strategy of antenna selection is beyond the scope of the thesis. On the other hand, all the users are assumed to use isotropic antennas.

3.2.3 Spectrum Model

The spectrum to be used is modelled following the LTE convention, which is implemented in current practical networks. Each BS in the system is permitted to use the 20 MHz system bandwidth with the carrier frequency of 3.5 GHz. The system bandwidth consists of 100 Physical Resource Blocks, each of which has 180 kHz bandwidth and is mapped to a Virtual Resource Block (VRB). The uplink transmission is considered in the research and the type 0 uplink resource allocation as standardised in [71] is chosen. Four consecutive VRBs selected as a spectrum bundle are assigned to a user during data transmission. Downlink transmission is not considered because it does not affect the key results investigated and the RRM is not the topic of the thesis. The choice of the uplink transmission enables the user transmission power to be directly captured and the downlink power is considered as introduced later.

3.2.4 Radio Propagation Model

The SCNs investigated are assumed to be deployed in urban areas such as stations and airports for capacity enhancement, so the WINNER II model B3 in [72] is used for the path loss calculation as

$$PL_{\mathrm{LOS}} = 13.9 \log_{10} d + 64.4 + 20 \log_{10}(\frac{f_c}{5}) + SF_{\mathrm{LOS}}, \tag{3.4a}$$

$$PL_{\mathrm{NLOS}} = 37.8 \log_{10} d + 36.5 + 23 \log_{10}(\frac{f_c}{5}) + SF_{\mathrm{NLOS}}, \tag{3.4b}$$

$$p_{\mathrm{LOS}} = \begin{cases} 1, & d \leq 10, \\ \exp\left(-\frac{d-10}{45}\right), & d > 10, \end{cases} \tag{3.4c}$$

where PL_{LOS} and PL_{NLOS} represent the path loss in dB of line-of-sight (LOS) and non-line-of-sight (NLOS) paths, respectively. d is the distance in metres between a transmitter and a receiver and f_c is the carrier frequency in GHz. SF_{LOS} and SF_{NLOS} are the log-normal shadowing fading variables for LOS and NLOS paths with the standard deviations of 3 dB and 4 dB, respectively. p_{LOS} denotes the probability of having a LOS path for a specific link.

3.2.5 Link Model

Practically, the channel capacity cannot reach the theoretical bound according to the Shannon-Hartley theorem. Hence, the performance of a connecting link is evaluated with the truncated

Shannon bound suggested in [73] in form of the throughput at the receiver end of a link. This actually reflects the transmission data rate with the unit, bps/Hz, which is given as

$$Throughput = \begin{cases} 0, & SINR < SINR_{\min}, \\ \alpha \log_2 (1 + SINR), & SINR_{\min} < SINR < SINR_{\max}, \\ Throughput_{\max}, & SINR > SINR_{\max}. \end{cases} \tag{3.5}$$

In the equation, the attenuation factor, α considers implementation losses. $SINR_{\min}$ and $SINR_{\max}$ are the lower and the upper threshold in dB where the throughput reaches the minimum and the maximum, respectively. $Throughput_{\max}$ is the maximum throughput achievable in bps/Hz. The parameters, α, $SINR_{\min}$ and $Throughput_{\max}$ can be chosen for different modem implementations and link conditions. Their values are determined to be 0.4, -10 dB and 2 bps/Hz for the uplink transmission as given in [73]. $SINR_{\max}$ can be derived from the equation based on $Throughput_{\max}$. $SINR_i$ ($i \in \mathbb{N}^*$) can be obtained by

$$SINR_i = \frac{S_i}{\sigma^2 + \sum I_i}. \tag{3.6}$$

Based on the assumption of a 5 dB noise figure at BSs, 300 K temperature and the whole system bandwidth, the noise floor σ^2 is set to -96 dBm. Hence, the SINR of channel i at a receiver can be calculated by dividing the receiving signal power of channel i by the sum of the noise floor and the total interference power of channel i at the receiver, $\sum I_i$, where S_i and I_i can be acquired by combining the antenna gains, the respective transmission power and the corresponding path losses.

3.2.6 Traffic Model

The traffic model considered in this thesis is to capture the data traffic pattern in wireless networks and to generate traffic for performance evaluation. Although the data traffic consists of packet streams, modelling in form of packets is very complex and time-consuming, which means a large number of arrival events to be simulated. For simplicity, the file model, as suggested in [74], is adopted to model the temporally varying data traffic offered by each user, where file arrivals conform to a HPPP. In a HPPP, the probability that the number of events (file arrivals), k ($k \in \mathbb{N}$), occurs in the time interval $(t, t + \tau]$ with the duration τ,

follows a Poisson distribution. This can be formulated as

$$\mathbf{P}\Big(N(t+\tau) - N(t) = k\Big) = \frac{\mathrm{e}^{-\lambda\tau}(\lambda\tau)^k}{k!}, \tag{3.7}$$

where $N(t)$ counts the number of the events by the time t. The mean interarrival time is equal to λ and the mean file arrival rate is therefore $1/\lambda$. Considering a special case of the HPPP with $k = 0$, the above equation then specifies the probability of no file arrivals in a period of time τ. Its complementary, the probability of having at least one file arrival in a period of time τ, can be derived as

$$\mathbf{P}\Big(N(t+\tau) - N(t) \geq 1\Big) = 1 - \frac{\mathrm{e}^{-\lambda\tau}(\lambda\tau)^0}{0!} = 1 - \mathrm{e}^{-\lambda\tau}. \tag{3.8}$$

In other words, the probability of having the interarrival time less than or equal to τ is equivalent to the above result by assuming the time between any two file arrivals is τ. This description refers to a cumulative distribution function (cdf) format and can be written as

$$F(\tau;\lambda) = \begin{cases} 1 - \mathrm{e}^{-\lambda\tau}, & \tau \geq 0, \\ 0, & \tau < 0. \end{cases} \tag{3.9}$$

According to $F(\tau;\lambda)$, the interarrival time between any two file arrivals conforms to an exponential distribution. As suggested in [74], all files have the size of 4 Mbits and the mean file arrival rate is varied to get different traffic densities. Traffic density is calculated by multiplying $1/\lambda$ by the file size and the number of users divided by the simulated area. It can be further normalised to the system bandwidth.

3.2.7 BS Power Consumption Model

Power consumption of a BS is required to be appropriately modelled for the evaluation of the overall power consumption of a SCN, which is one of the main metrics to assess the effectiveness of sleep mode operation algorithms investigated in the thesis. Two BS power consumption models are provided based on the types of the BSs for two architectures studied in the research. In the distributed LTE architecture, eNBs are the access nodes under consideration while RRHs are the counterparts in a C-RAN architecture.

Table 3.1: Assumptions of the transient states

Level of Energy Consumption	$P_{\text{tran}}^{\text{on}}$	$T_{\text{tran}}^{\text{on}}$	$P_{\text{tran}}^{\text{off}}$	$T_{\text{tran}}^{\text{off}}$
High	$2P_0$	2 s	P_0	0.2 s
Medium	$2P_0$	1 s	P_0	0.1 s
None	$2P_0$	0 s	P_0	0 s

eNB Power Consumption Model

For the eNB power consumption, the BS power model in [7] is adopted, which is derived based on measured data. As a single transceiver chain is assumed, the instantaneous power consumption of an active eNB, P_{B}, can be simplified to

$$P_{\text{B}} = P_0 + \Delta_{\text{p}} P_{\text{out}}, \tag{3.10}$$

where P_0 is the static power of the active mode. Δ_{p} is the slope of the load-dependent power consumption, P_{out}, which is assumed linear to the eNB load and reaches $P_{\max}$ at the maximum load. For a system with only uplink simulated, the downlink load is assumed identical to the uplink one. P_{S} is used to denote power consumption of the sleep mode, which is a constant number smaller than P_0 with some hardware modules deactivated. The power model here is an approximation of the power consumption of all hardware modules, e.g. power amplifiers, radio fronts, baseband processing, DC-DC and main supply. The data of femtocell base stations in [7] is used (P_0 = 4.8 W, P_{S} = 2.9 W, Δ_{p} = 8.0, $P_{\max}$ = 0.05 W), for the SCNs investigated.

The energy consumed during transient states can also been considered. The transient states are short-period states before an eNB is completely switched on or completely switched to a sleep mode. The switch-on transient state is modelled by the constant power consumption $P_{\text{tran}}^{\text{on}}$ and the total time needed to complete the switch-on transition $T_{\text{tran}}^{\text{on}}$. On the other hand, the transient state of turning an eNB to a sleep mode can be modelled by the constant power consumption $P_{\text{tran}}^{\text{off}}$ and the required time $T_{\text{tran}}^{\text{off}}$ [75]. As the exact values are not provided, different reasonable values can be assumed to study the effect of transient states, which are classified into high, medium and none energy consumption transient state as shown in Table 3.1.

Table 3.2: Related parameter values of the RRH power model

N_{ant}	2	BW	40 MHz	P_{RF}	0.8 W
$P_{\text{PA-max}}$	0.8 W	σ_{DC}	6.4%	σ_{MC}	7.7%

RRH Power Consumption Model

The RRH power consumption model developed in [76] is adopted. This model is based on [7], but specifically considers the linear power consumption of power amplifiers, which is a more advanced technology as introduced in Section 2.4.1. The RRH case is also particularly considered with reasonable data given. The RRH instantaneous power depends on the RRH state, either the active mode or the sleep mode. The active RRH power consumption P_B as given in [76], is modelled as

$$P_B = \frac{N_{\text{ant}} \frac{BW}{10} P_{\text{RF}} + L_b P_{\text{PA-max}}}{(1 - \sigma_{\text{DC}})(1 - \sigma_{\text{MC}})}, \tag{3.11}$$

where N_{ant} and BW in MHz stand for the number of antennas used by a RRH and the system bandwidth, respectively. P_{RF} in Watts and $P_{\text{PA-max}}$ in Watts denote the RF base consumption and the maximum PA transmission power, respectively. σ_{DC} and σ_{MC} are the main supply losses factor and the DC-DC conversion losses factor, respectively. L_b is the percentage load of a RRH. Realistic values are given in Table 3.2.

When a RRH goes into the sleep mode, the RF module and the PA module of the RRH do not need to be active, and are assumed able to be turned off instantly. The instantaneous power of a sleeping RRH P_S thus falls to zero based on Equation (3.11) because the model given in [76] does not consider the power consumption such as from the backhaul connection and the basic processing unit. Therefore, P_S is modelled according to [77], which is 15% of P_B at zero loads. Since only the RF and PA modules have to be deactivated for a sleep mode, $T_{\text{tran}}^{\text{off}}$ and $P_{\text{tran}}^{\text{off}}$ = are assumed zero to model a rapid deactivation. However, it is suggested that the wake-up transient state is considered not to be negligible. The wake-up time $T_{\text{tran}}^{\text{on}}$ is chosen to be 0.5 seconds and $P_{\text{tran}}^{\text{on}} = 0.5P_B$ as given in [77].

3.3 Simulation Techniques

This section describes the features of the simulation tool, MATLAB, used in the research and presents the information about how the simulation is carried out to investigate the

performance of the SCNs at a system level.

3.3.1 Simulation Tools

With respect to the simulation tools for the performance evaluation of wireless systems, there is a broad range of options available for different layers of a system and different purposes. Among all the candidates, MATLAB is chosen as a programming platform to carry out the system simulation. MATLAB is developed by Mathworks as a commercial software and provides powerful capability of matrix computation and graphic handling tools. Through the past years, MATLAB has accumulated much popularity among engineers, scientists and other professionals. Different from programming languages like C, C++ or JAVA, MATLAB does not perform very well in execution efficiency as those low level languages, which map efficiently to machine instructions. However, MATLAB provides an excellent visible programming environment with accessible data values in the workspace window, which is convenient to view. It also contains easy-to-use debugging tools to find logical errors. The syntax of MATLAB is extremely simple, which makes it more efficient to program and saves the time of programmers. The default data type in MATLAB is double float, which is very suitable for numerical computation without the need to define explicitly. Matrix manipulation in MATLAB is very powerful and easy to use with simple syntax. The operation of matrix computation is highly optimised by the software developers and fits the research, which contains the system level simulation with a large amount of data to be processed. As long as a programmer is familiar with this the matrix operation, much time can be saved in programming and executing the codes.

3.3.2 System Level Simulation

In the system level simulation, the high level performance measurement of the wireless system should be taken to validate the effectiveness of different schemes. Fig. 3.3 shows the structure of the simulator developed for the research proposed in the thesis. As indicated from the figure, the simulation adopts the system models to generate measurements for performance evaluation. To obtain a large amount of reliable statistical measurements for the final analysis, sufficient realisations of the wireless system should be fed into the simulator to reduce the effect of random variations in the system models. For the efficiency of running the codes in MATLAB, the wireless system is converted to pre-generated data matrices based on the given

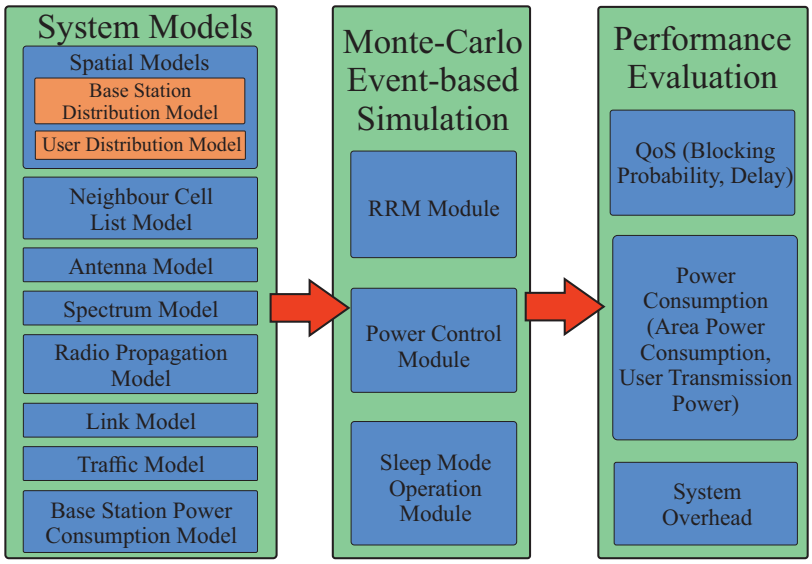


Figure 3.3: Structure of the simulator.

models before running the simulator. The random values in the models are all generated with pseudo-random numbers, which are typically generated by algorithms involving deterministic processes and conform to statistical distributions. In MATLAB, they are generated by using pseudo-random number generators.

When having the digitised wireless system on hand, a simulation using the Monte-Carlo method can be carried out to get the measurements. The Monte-Carlo method adopts repeated random sampling to get numerical results. With the pre-generated data matrices as sampled data of the modelled practical wireless systems, the simulation can be run repeatedly to get the aggregated measurements of the metrics of interest. The simulation takes an event-based strategy meaning that the simulator only responds to an event occurrence instead of updating system conditions at discrete time points. The wireless system does not change during any two events, where an event can be a service request from a user, completion of a file transmission, a network adjustment, etc. This approach can significantly reduce the simulation time consumed compared with a time-based strategy, which updates a system even at the time when system conditions do not change.

In the RRM module, users are assumed to acquire radio resources from the nearest active BSs in terms of path loss and the spectrum resource bundle with the highest SINR is allocated, which ensures the best link a BS can provide for the time when it requires the data service. A user is blocked to avoid congestion and low data transmission rate in the system if no resource bundle has SINR above 5 dB. For simplicity, after the successful assignment, the user is assumed to occupy the resource bundle from the associated BS until the completion of

the file transmission even if the data rate is zero temporarily during the transmission. In this case, file dropping and retransmission are not considered by assuming continuous transmission on the breakpoint. In the power control module, open loop power control with signal-to-noise ratio (SNR) of 25 dB is applied for the uplink data transmission to adjust the transmission power. The value of the target SNR is obtained by simulation, where the peak throughput can almost been achieved with 25 dB SNR. A specific sleep mode operation algorithm is applied in the sleep mode operation module. As handover during data transmission is not explicitly considered, when an BS is to be switched to a sleep mode, it waits until the end of the data transmission and stops admitting future users. During this state, power consumption of an active mode with a zero load should be used to model the instantaneous power of a BS.

In the simulation, the measurement is only taken after a specified period of time has elapsed and the wireless system is relatively stable. The measurements of interest are aggregated after a simulation for performance evaluation, which is introduced in detail in the next section.

3.4 Performance Evaluation

With a large number of measurements produced after a simulation, system performance for different schemes can be quantified. The performance evaluation methods presented in this section is particularly needed for evaluating the effectiveness of sleep mode operation algorithms, which is one of the main topics studied in the thesis.

3.4.1 QoS Evaluation

QoS is one of the conventional metrics to measure the performance of a communication network and it is critical in the estimation of the capacity provided by a small cell system. From another perspective, the effectiveness of a sleep mode operation algorithm is that it can maintain the same QoS delivered to the end users as the system without sleep modes when reducing the overall network power consumption. Therefore, QoS should be evaluated, which can be represented by the combination of blocking probability and delay.

The blocking probability for the file arrival traffic model adopted in the research can be defined as the ratio of the total file transmission requests blocked and the total file arrivals during a simulation. As file transmission interruptions are not considered, the average

blocking probability of a wireless system $\overline{BP}$ can be formulated as

$$\overline{BP} = \frac{N_{\text{blk}}}{N_{\text{tot}}} \times 100\%, \tag{3.12}$$

where N_{tot} is the total file arrivals and N_{blk} is the total file transmission requests blocked.

With respect to delay, it represents the time required for a receiver to receive all the data of a file via the wireless link. With a fixed file size, delay directly reflects the data rate of a file transmission and the quality of the allocated frequency resource in terms of SINR. The average delay $\overline{\varepsilon}$ is formulated as

$$\overline{\varepsilon} = \frac{\sum_{i=1}^{N_{\text{tot}}-N_{\text{blk}}} \varepsilon_i}{N_{\text{tot}} - N_{\text{blk}}}, \tag{3.13}$$

where ε_i ($i \in \mathbb{N}^*$) is delay of the transmission of file i, which is counted from the time when the first bit is sent from a transmitter until the time when the last bit is received by a receiver.

3.4.2 Power Consumption Evaluation

When discussing energy efficiency investigated in the thesis, there are many metrics to quantify the performance of this aspect. One approach is relating the energy consumption to the throughput of a network, yielding bits/Joule as the unit. Another simple method is to directly use the overall power consumption of a network. The total network power consumption can be normalised to the power consumption with all BSs at active modes if the minor load-dependent power consumption is neglected. In the thesis, area power consumption is mainly considered, where the total network power consumption is normalised by the total area of the network. This method is chosen due to the SCNs studied in this thesis, whose system performance is usually considered based on a unit area. Hence, with power consumption of a BS modelled, the equivalent instantaneous area power consumption of a network, $\widehat{P_{\text{t}}^{N}}$ considered in a period of time τ, can be calculated as

$$\widehat{P_t^{\text{N}}} = \frac{\int_0^\tau \sum_{b=1}^{|\mathcal{B}|} P_B \mathbf{1}_{\mathcal{B}^{\text{on}}}(B_b) + P_S \mathbf{1}_{\mathcal{B}^{\text{off}}}(B_b) \mathrm{d}t + n_{\text{on}} P_{\text{tran}}^{\text{on}} T_{\text{tran}}^{\text{on}} + n_{\text{off}} P_{\text{tran}}^{\text{off}} T_{\text{tran}}^{\text{off}}}{|\mathcal{A}|\tau}, \tag{3.14}$$

where $\mathbf{1}_{\mathcal{B}}(\cdot)$ is an indicator function of $\mathcal{B}$, which returns 1 if the variable is a member of $\mathcal{B}$ or returns 0 otherwise. The BS set $\mathcal{B} = \{B_b\}$ ($b \in \mathbb{N}^*$) is the union of active BSs $\mathcal{B}^{\text{on}}$ and sleeping BSs $\mathcal{B}^{\text{off}}$. $|\mathcal{B}|$ is the total number of the BSs in a network. n_{on} and n_{off} denote the times of switching on and switching to sleep modes in the network during τ. Other parameters are defined as aforementioned.

The average transmission power of users, $\widehat{P^{\mathrm{u}}}$ is also investigated to evaluate the influence on users from the application of sleep modes and is defined as

$$\widehat{P^{\mathrm{u}}} = \frac{\sum_{m=1}^{M} \sum_{f=1}^{F_m} P_{m,f} t_{m,f}}{\sum_{m=1}^{M} \sum_{f=1}^{F_m} t_{m,f}} \tag{3.15}$$

where $P_{m,f}$ is the transmission power of user m during the transmission of the file f and $t_{m,f}$ is its corresponding elapsed time during file transmission, which is ε_f of user m. F_m is the number of files transmitted by user$_m$ during the simulation and M is the number of users in the network.

3.4.3 System Overhead Evaluation

The application of a sleep mode may yield extra system overheads, which occupy additional computational resources and may consume redundant power, which should be investigated. In sleep mode operation, the overhead contributed by the switch on-offs of BSs is unavoidable, however, can be minimised to keep a relatively stable network. To quantify this metric, the frequency of switch on-offs can be used with a unit of times per unit time. On the other hand, if a sleep mode operation algorithm requires coordination and cooperation among BSs, a certain level of information exchange is needed. It is important especially in a distributed architecture, where information exchange occupies extra radio resources for control signals. This metric can be quantified by combining together the frequency of information exchange and the sizes of information messages. However, it is hard to determine the message sizes without knowing the actual protocol. In the thesis, only the frequency of information exchange is considered with a unit of times per unit time.

3.5 Summary

This chapter presents the methodology used to study the sleep mode operation of SCNs, forming the basis of the research conducted. Both ideal and non-ideal spatial distributions are considered for different SCN scenarios. Power consumption models dedicated to small cell BSs are presented for the distributed and the centralised architecture, respectively. Other related system models are introduced. The Monte-Carlo event-based simulation is adopted. In the simulation, measurements of interest are taken for the key performance evaluations of sleep mode operation including QoS, power consumption and system overheads.

Chapter 4

Distance Distributions

Contents

4.1 Introduction

Since the deployment of conventional cellular networks, grid-based models (e.g. the hexagonal grid model) have been extensively used to model the distribution of macro BSs. However, the locations of BSs in reality are observed to have irregular patterns due to the existence of hotspot areas and constraints of actual site deployments. In the SCNs investigated in the thesis, small cell BSs impose more obvious deviations on the BS locations from idealised

grid models, where the randomness can be captured by adopting stochastic geometry models [27]. Stochastic geometry has been applied in modelling ad hoc networks for decades. As a tractable mathematical tool for theoretical analysis, it has also gained much interest in the performance analysis and scheme comparison for cellular networks in the recent research [27] as it provides a better approximation for the networks in reality.

Among all stochastic models, the homogeneous Poisson Point Process (HPPP) has gained the highest popularity because of its simplicity and mathematical tractability for the sake of analysing and characterising network performance. Although PPP brings about abundant closed-form results, the PPP is not a good model in the context of cellular networks [27, 28]. Since there is cell planning in the deployment of BSs, a repulsive point process should be considered to prevent any two BSs being arbitrarily close to each other. This requirement can be satisfied by introducing a minimum separation distance or a hard-core distance in a secondary point process beyond the fundamental HPPP, which yields a Matérn Hard-core Point Process (MHPP) [78]. The MHPP has two types, type II of which has gained a higher interest in modelling wireless networks [27, 28, 34, 79] and is selected to model the SCNs as introduced in Chapter 3.

Due to the fact that path loss between a transmitter and a receiver is dependent upon the distance between them, any analysis of power-related metrics (or simply distances themselves) in cellular systems of random layouts ultimately falls into the quantification of the distance distributions. These related system metrics are essential in optimal scheme exploration or analysis of system performance, e.g. transmission power consumption, outage probability. For BSs modelled by stochastic processes like MHPP, the distances under consideration include the contact distance and the nearest-neighbour distance [30]. The contact distance is the distance between a generic point and the nearest point generated by a point process, corresponding to the distance between a user and a BS in a cellular network. Whereas the nearest-neighbour distance refers to the distance from a point to its nearest neighbour, both of which are generated by a point process, such as the distance from a BS to its nearest BS. The generalisation to the distance to the nth ($n \geq 2$) nearest point enables more complex calculation such as interference. Therefore, determined distance probability distributions make mathematical analysis possible to evaluate the global system performance based on stochastic models, which is about sleep mode operation for on-demand radio resource provisioning and its related metrics within the scope of the thesis.

In the following chapters, type II MHPP will be extensively used for modelling the SCNs

with non-ideal distributions. However, the analysis of a MHPP is limited by the lack of a probability generation function [65], which is indispensable in the derivation of distance probability distributions. The known approaches used in the literature to approximate the distance probability distributions of a MHPP exploit the corresponding functions of its parent HPPP and replace the intensity, λ_p, with other intensity distributions, such as [80] for the contact distance distribution and [33] for the nearest-neighbour distance distribution. However, the existing methods underestimate the distance probability distributions, which may result in coarse modelling and inaccurate further analysis. Meanwhile, the distances to the nth ($n \geq 2$) nearest point have not been considered.

As the main contribution, this chapter proposes derivations and approximations of the distance distributions of MHPP for the rest of the research proposed in the thesis as well as for abundant conventional and potential applications. To understand the type II MHPP well,the distance distributions of its parent HPPP are introduced in Section 4.2 and the properties of the type II MHPP are presented in Section 4.3. They indicate the mathematical intractability of the type II MHPP and form the basis of the approximation approaches proposed in Section 4.4.

4.2 Distance Distributions of HPPP

HPPP is the parent point process of type II MHPP. Considering a 2D case, $B(x,R)$ denotes a closed disc in the 2-dimensional Euclidean space $\mathbb{R}^2$, centred at x with radius of R. For a HPPP $\Phi_p \subset \mathbb{R}^2$ with intensity λ_p, the probability of having k points in $B(o,r)$ can be derived from the general probability generation function [30] introduced in Chapter 2 and can be written as

$$\mathbf{P}\Big(\Phi_p\big(B(o,r)\big) = k\Big) = \exp\Big(-\lambda_p \pi r^2\Big)\frac{\big(\lambda_p \pi r^2\big)^k}{k!}. \tag{4.1}$$

The function can be applied at an arbitrary location $u \in \mathbb{R}^2$ including the origin o since the HPPP is stationary and isotropic. The distance distributions can be derived based on Equation (4.1). The introduction of the distance distributions of HPPP provides the background information and the recognition of the tractability of HPPP, which indicates the difficulties in deriving the distance distributions of MHPP. The distance distributions of HPPP will also be used for the comparison with the MHPP in Section 4.4.

4.2.1 Contact Distance Distribution

The contact distance distribution $F(r)$ in terms of the cdf of a HPPP Φ_{p} at a generic point $u \in \mathbb{R}^2$ is the distance between u and the nearest point generated by the HPPP [29], which can be formally defined as

$$F(r) = \mathbf{P}(\min_{x_i \in \Phi_{\mathrm{p}}} \|u - x_i\| \leq r), \tag{4.2}$$

where $\|\cdot\|$ is the Euclidean distance and $i \in \mathbb{N}^*$. In other words, $F(r)$ is the probability of having at least one point $x_i \in \Phi_{\mathrm{p}}$ in a disc of radius r centred at u since r is equal to or larger than the distance between u and its nearest point $x_i \in \Phi_{\mathrm{p}}$. With Equation (4.1), it continues the derivation as

$$F(r) = \mathbf{P}\Big(\Phi_{\mathrm{p}}\big(B(u,r)\big) \geq 1\Big) = 1 - \mathbf{P}\Big(\Phi_{\mathrm{p}}\big(B(u,r)\big) = 0\Big) = 1 - \exp\big(-\lambda_{\mathrm{p}}\pi r^2\big). \tag{4.3}$$

Due to the stationary property of the HPPP, $F(r)$ does not depend on the location of u. For SCNs modelled by HPPPs, $F(r)$ corresponds to the distribution of the distance from an arbitrary user to its nearest BS.

4.2.2 Nearest-neighbour Distance Distribution

The nearest-neighbour distance distribution $G(r)$ in terms of the cdf of a HPPP Φ_{p} at a target point $x_i \in \Phi_{\mathrm{p}}$ refers to the distance from x_i to its nearest neighbour in Φ_{p} [29], which can be formally defined as

$$G(r) = \mathbf{P}(\min_{x_j \in \Phi_{\mathrm{p}} \setminus x_i} \|x_i - x_j\| \leq r), \tag{4.4}$$

where $i \neq j$ ($i,\ j \in \mathbb{N}^*$). Similarly, $G(r)$ is the probability of having at least two points of Φ_{p} (including x_i itself) in $B(x_i,r)$. Consider x_i at o, $G(r)$ can be derived according to the property of the reduced Palm distribution of the HPPP as

$$\begin{aligned} G(r) &= \mathbf{P}\Big(\Phi_{\mathrm{p}}\big(B(o,r)\big) \geq 2 \| x_i\Big) = \mathbf{P}\Big(\Phi_{\mathrm{p}}\big(B(o,r) \setminus \{x_i\}\big) \geq 1 \| x_i\Big) = \mathbf{P}\Big(\Phi_{\mathrm{p}}\big(B(o,r)\big) \geq 1\Big) \\ &= 1 - \mathbf{P}\Big(\Phi_{\mathrm{p}}\big(B(o,r)\big) = 0\Big) = 1 - \exp\big(-\lambda_{\mathrm{p}}\pi r^2\big). \end{aligned} \tag{4.5}$$

The theory of the reduced Palm distribution and the related notations can be found in Appendix A. As can be seen, $G(r)$ has the same result as $F(r)$ meaning that $G(r)$ and $F(r)$ of a HPPP are equivalent. Like $F(r)$, $G(r)$ does not depend on the location of x_i under

consideration. For SCNs modelled by HPPPs, $G(r)$ corresponds to the distribution of the distance from an arbitrary BS to its nearest BS.

4.2.3 nth Distance Distribution

The nth distance distribution considers the distance between an arbitrary point $u \in \mathbb{R}^2$ and its nth nearest point or the distance between $x_i \in \Phi_\mathrm{p}$ and its nth nearest point in Φ_p, denoted by $F_n(r)$ or $G_n(r)$, respectively. Being the probability that the distance from u to its nth nearest point is not larger than r, $F_n(r)$ is calculated by equalling to having at least n points in $B(u,r)$ as

$$F_n(r) = \mathbf{P}\Big(\Phi_\mathrm{p}(B(u,r)) \geq n\Big) = 1 - \mathbf{P}\Big(\Phi_\mathrm{p}(B(u,r)) \leq n-1\Big). \tag{4.6}$$

On the other hand, the distance distribution between $x_i \in \Phi_\mathrm{p}$ and its nth nearest-neighbour point in Φ_p is represented by its cdf $G_n(r)$. $G_n(r)$ is the probability that the distance from x_i to its nth nearest neighbour point is not larger than r, meaning that there are at least $n+1$ points (including x_i itself) in $B(x_i,r)$. Based on Equation (4.1) and (4.5), it is straightforward to derive

$$F_n(r) = G_n(r) = 1 - \exp(-\lambda_\mathrm{p}\pi r^2) \sum_{k=0}^{n-1} \frac{(\lambda_\mathrm{p}\pi r^2)^k}{k!}. \tag{4.7}$$

A generalised 3D case can be found in [29,30]. For SCNs modelled by HPPPs, $F_n(r)$ or $G_n(r)$ corresponds to the distribution of the distance from an arbitrary user or BS to its nth nearest BS.

4.3 Study of Type II MHPP

This section introduces some important fundamental results of type II MHPP, which will be used in the following section. The realisation of a MHPP in simulation is also presented for the evaluation of either MHPP itself or a SCN modelled by a MHPP.

Considering a parent HPPP Φ_p with intensity λ_p, all the primary points $x_i \in \Phi_\mathrm{p}$ ($i \in \mathbb{N}^*$) are associated with random marks t_i of independent uniform distributions over the range $(0,1)$. To render a type II MHPP Φ_m of the hard-core distance δ, each x_i is deleted if there is another primary point with a smaller mark in $B(x_i,\delta)$.

4.3.1 Retaining Probability and Pairwise Interaction

The retaining probability p of a typical point of Φ_{p} can be defined and derived as

$$p = \mathbf{P}(x_i \in \Phi_{\mathrm{m}} \mid x_i \in \Phi_{\mathrm{p}}) = \frac{1 - \exp(-\lambda_{\mathrm{p}}\pi\delta^2)}{\lambda_{\mathrm{p}}\pi\delta^2}. \tag{4.8}$$

The results are given in [29, 30] without the derivation provided. The derivation in detail is provided in Appendix B. It is obvious that the retaining probability of a primary point without conditioning on other primary points is identical, from which the final intensity of Φ_{m}, λ_{m}, is $\lambda_{\mathrm{p}}p$. Intuitively, due to the convergence of the numerator of $\lambda_{\mathrm{p}}p$, $\lambda_{\mathrm{m}} < 1/(\pi\delta^2)$. For extraordinary large λ_{p}, λ_{m} is mainly determined by δ.

The probability of retaining two points x_i and x_j $(j \in \mathbb{N}^*, j \neq i)$ is denoted as $k(s)$ and can be formulated as

$$k(s) = \mathbf{P}(x_i \in \Phi_{\mathrm{m}}, x_j \in \Phi_{\mathrm{m}} \mid x_i \in \Phi_{\mathrm{p}}, x_j \in \Phi_{\mathrm{p}}). \tag{4.9}$$

$k(s)$ indicates the pairwise interaction between any two primary points. It is a function of the Euclidean distance between them, $s = \|x_i - x_j\|$. For $s \leq \delta$, $k(s) = 0$ because the hard-core area of a point repels the other. While for $s > \delta$, $k(s)$ is given by [29]

$$k(s) = \frac{2V_\delta(s)(1 - \mathrm{e}^{-\lambda_{\mathrm{p}}\pi\delta^2}) - 2\pi\delta^2(1 - \mathrm{e}^{-\lambda_{\mathrm{p}}V_\delta(s)})}{\lambda_{\mathrm{p}}^2 V_\delta(s)\pi\delta^2(V_\delta(s) - \pi\delta^2)}, \tag{4.10}$$

where $V_\delta(s) = v_2(B(x_i,\delta) \cup B(x_j,\delta))$ represents the area of the union of the two circles of radii δ centred at x_i and x_j. Since it is hard to find the detailed derivation in the literature, it is provided in Appendix C. For $2\delta \geq s \geq 0$,

$$V_\delta(s) = 2\pi\delta^2 - 2\delta^2 \arccos(\frac{s}{2\delta}) + s\sqrt{\delta^2 - \frac{s^2}{4}}. \tag{4.11}$$

$V_\delta(s) = 2\pi\delta^2$ if $s > 2\delta$ as $B(x_i,\delta) \cap B(x_j,\delta)$ vanishes. $k(s)$ indicates the correlation between two primary points in the thinning process, which will be used in the later analysis.

4.3.2 Simulation of a MHPP

In the dependent thinning process, the retaining of a primary point is affected by its near primary points. Considering a MHPP over $\mathbb{R}^2$, each primary point has identical probability conditions including the probability of having a certain number of primary points in its hard-

core area. However, there is an edge when a realisation of a MHPP on a finite subset $W \subset \mathbb{R}^2$ is considered. In general, the primary points at the edge have fewer primary points in their hard-core areas. According to Equation (B.1), these primary points have higher retaining probabilities than the primary points far away from the edge. Hence, the edge primary points can affect the overall probability characteristics of the realisation of the MHPP on W.

To avoid the circumstance, a periodic edge-correlation technique [29] is adopted for the mitigation of the edge effect. After the generation of the parent HPPP of a MHPP on a square W and the random marks of all the primary points, W including all the primary points is replicated to 8 copies, which are shifted around the central W to constitute a large square of 9 identical grids. The thinning process is applied to the original primary points and the replicated ones and only the primary points in W are considered as the realisation of the MHPP. With the periodic edge-correlation, the edge primary points are given the identical retaining probability as the other ones. W has to be large enough compared with δ to exclude a primary point itself in its thinning process.

4.4 Distance Distributions of Type II MHPP

The derivation of distance distributions for HPPP and the properties of MHPP have been investigated in the previous sections. They form the fundamental of the analysis regarding the nth distance distributions for MHPP, which is the main contribution of this section. For the $n = 1$ case, the distance distributions refer to the contact distance distribution $F(r)$ and the nearest-neighbour distance distribution $G(r)$. In the practical SCNs modelled by MHPPs, they reflect the distance from either a user or a BS to the nearest BS. In the wireless environment, the nearest distance usually corresponds to the strongest signal, which can be considered particularly. With the generalisation to the $n \geq 2$ case, the distance distributions indicate a probability of having a certain distance from either a user or a BS to any specific BS in a SCN. Some of the results will be directly used in Chapter 5 and Chapter 7.

4.4.1 Contact Distance Distribution

The contact distance distribution $F(r)$ of a MHPP Φ_{m} at an arbitrary location $u \in \mathbb{R}^2$ is formally defined as

$$F(r) = \mathbf{P}(\min_{x_i \in \Phi_{\mathrm{m}}} \|u - x_i\| \leq r). \tag{4.12}$$

Similarly as the one of a HPPP, it is the probability of having at least one point in a disc of radius r. Since the MHPP is also stationary and isotropic, $F(r)$ does not depend on the location of u.

A closed-form $F(r)$ can be rigorously derived for $\delta/2 \geq r \geq 0$. In this case, $F(r)$ is effectively the probability of having one point in the disc $B(u,r)$ since any two primary points in $B(u,r)$ (if any) have a pairwise distance less than or equal to δ, which forces one of the points to be deleted in the thinning process after Φ_{p}. Assuming the solid dot to be retained as shown in the example given in Fig. 4.1 (a), its hard-core area covers the disc centred at the target arbitrary point u and prevents any other primary points in the disc from succeeding in the thinning process. Therefore, $F(r)$ can be derived as

$$\begin{aligned} F(r) =& \mathbf{P}\Big(\Phi_{\mathrm{m}}(B(u,r)) \geq 1\Big) \\ =& \mathbf{P}\Big(\Phi_{\mathrm{m}}(B(u,r)) = 1\Big) \\ =& \sum_{a=1}^{\infty} \mathbf{P}\Big(\Phi_{\mathrm{p}}(B(u,r)) = a\Big)\binom{a}{1} \\ & \cdot \sum_{b=0}^{\infty} \mathbf{P}\Big(\Phi_{\mathrm{p}}(B(x_{\mathrm{c}},\delta) \setminus B(u,r)) = b\Big) \\ & \cdot \mathbf{P}\Big(\Phi_{\mathrm{p}} \cap B(x_{\mathrm{c}},\delta) \setminus x_c \cap \Phi_{\mathrm{m}} = \varnothing\Big), \end{aligned} \tag{4.13a}$$

where x_{c} denotes a possible primary point that can be retained. As shown in Fig. 4.1 (a), it is only possible to retain x_{c} in $B(u,r)$ due to the hard-core area of x_{c} of radius δ (the circle depicted with the dashed line). It gives $\mathbf{P}\Big(\Phi_{\mathrm{m}}(B(u,r)) = 1\Big)$ in Equation (4.13a). Then for each possibility of generating a primary points in $B(u,r)$ before the thinning process, $\binom{a}{1}$ indicates a alternative combinations (with one element in the combination) when choosing x_{c} to be retained from a generated primary points. The following term sums up the probabilities of having other b primary points in the rest of $B(x_{\mathrm{c}},\delta)$, which is $B(x_{\mathrm{c}},\delta) \setminus B(u,r)$. Then, deleting all the primary points excluding x_{c} in $B(x_{\mathrm{c}},\delta)$ is considered to retain only x_{c} for computing $\mathbf{P}\Big(\Phi_{\mathrm{m}}(B(u,r)) = 1\Big)$. The derivation continues as

$$
\begin{aligned}
F(r) = &\int_0^1 \sum_{a=1}^{\infty} \frac{(\lambda_\mathrm{p}\pi r^2)^a}{a!} \exp(-\lambda_\mathrm{p}\pi r^2) a \sum_{b=0}^{\infty} \frac{\left(\lambda_\mathrm{p}\pi(\delta^2 - r^2)\right)^b}{b!} \\
&\cdot \exp\left(-\lambda_\mathrm{p}\pi(\delta^2 - r^2)\right)(1-t_\mathrm{c})^{a+b-1}\mathrm{d}t_\mathrm{c} \\
= &\int_0^1 \lambda_\mathrm{p}\pi r^2 \exp(-\lambda_\mathrm{p}\pi\delta^2) \sum_{a=1}^{\infty} \frac{(\lambda_\mathrm{p}\pi r^2)^{a-1}}{(a-1)!}(1-t_\mathrm{c})^{a-1} \\
&\cdot \sum_{b=0}^{\infty} \frac{\left(\lambda_\mathrm{p}\pi(\delta^2 - r^2)\right)^b}{b!}(1-t_\mathrm{c})^b \mathrm{d}t_\mathrm{c} \\
= &\int_0^1 \lambda_\mathrm{p}\pi r^2 \exp(-\lambda_\mathrm{p}\pi\delta^2) \exp\left(\lambda_\mathrm{p}\pi r^2(1-t_\mathrm{c})\right) \\
&\cdot \exp\left(\lambda_\mathrm{p}\pi(\delta^2 - r^2)(1-t_\mathrm{c})\right)\mathrm{d}t_\mathrm{c} \\
= &\int_0^1 \lambda_\mathrm{p}\pi r^2 \exp(-t_\mathrm{c}\lambda_\mathrm{p}\pi\delta^2)\mathrm{d}t_\mathrm{c} \\
= &p\lambda_\mathrm{p}\pi r^2,
\end{aligned}
\tag{4.13b}
$$

where t_c is the random mark associated with x_c. The probability generation function of Φ_p is adopted to calculate the probability of generating a primary points before handling the retaining probability. Due to the fact that a random mark associated with a primary point is independently and uniformly distributed over the range $(0,1)$, the probability of deleting all the other primary points in $B(x_\mathrm{c},\delta)$ excluding x_c for a given t_c is calculated by multiplying $1-t_\mathrm{c}$ for $a+b-1$ times, which is the number of the other generated primary points in $B(x_\mathrm{c},\delta)$. The corresponding derivation in Equation (4.13b) also shows that the retaining probability of any of the a alternatives in this case is independent and is exactly p. The characteristic of an exponential function is utilised to convert from the power series. The above derivation is given in full for the purpose of completeness and the sake of explaining the following content. Finally in the result for $\delta/2 \geq r \geq 0$, $F(r)$ of Φ_m is determined by λ_p and δ, and is linear to the squared distance, r^2. It also signifies that $F(r)$ is equal to the expected number of points of Φ_m in $B(u,r)$ for $\delta/2 \geq r \geq 0$.

When dealing with the range of $r > \delta/2$, the flaw of the MHPP emerges in the derivation with the multiple retaining points. In $B(u,r)$, more than one primary point can be possibly retained, making it likely to have some to-be-retained points (the points considered in the combinations) with overlapping hard-core areas. $B(u,r)$ can only contain two retained points for $\sqrt{3}\delta/3 \geq r > \delta/2$, where $\sqrt{3}\delta/3$ can be obtained by considering the radius of a circumscribed circle of an equilateral triangle of side length equal to δ. In this case, $F(r) = \mathbf{P}\left(\Phi_\mathrm{m}(B(u,r)) = 1\right) + \mathbf{P}\left(\Phi_\mathrm{m}(B(u,r)) = 2\right)$. The largest difficulty appears in the calcu-

lation of the second term. In this scenario, $k(s)$ can be adopted to consider the probability of retaining these two points (instead of simply squaring p because of the interaction between near points) after selecting one of $\binom{a}{2}$ combinations (for $a \geq 2$) from the primary points. However, on reaching this, the random locations of these two points require two integrals over the disc $B(u,r)$, which yield complex computation and intractability. Moreover, this is just the case for $\mathbf{P}\big(\Phi_{\mathrm{m}}(B(u,r)) = 2\big)$ when $\sqrt{3}\delta/3 \geq r > \delta/2$. For the case where a bigger r needs to be considered to allow more possible points to be retained, the calculation includes more possible combinations with an extra summation required whose terms have similar forms to the whole Equation (4.13a). It also involves non-overlapping or multiple overlapping hard-core areas, where interactions among multiple to-be-retained points remain unknown. As shown in the example in Fig. 4.1 (b), the primary points deleted by one of the to-be-retained points, e.g. $x_{\mathrm{c}1}$, are likely to have larger marks and their existence in the overlapping areas increases the retaining probability of other to-be-retained points with overlapping areas, e.g. $x_{\mathrm{c}2}$ and $x_{\mathrm{c}3}$. Adding up the difficulties, it is not possible to derive a closed-form $F(r)$. Therefore, it is only feasible to approximate $F(r)$ to satisfy the demand of a straightforward distance distribution function for further analysis.

The $F(r)$ approximation is required to be continuous over the full range $(0,\infty)$, meanwhile meeting the requirements that $F(0) = 0$ and $\lim_{r\to\infty} F(r) = 1$. $F(r)$ has to be a straightforward, effective and accurate closed form for analytical tractability. λ_{p} and δ should be contained because they characterise different MHPPs. Specifically, the $F(r)$ approximation of a MHPP should approach the counterpart of its parent HPPP as $\delta \to 0$ indicating no thinning process. As a cdf, $F(r)$ also needs to have no units even when arguments of $F(r)$ have units in practical applications. The proposed model is obtained by heuristic approximation satisfying these conditions. It is shown as

$$F(r) = 1 - \exp(-qp\lambda_{\mathrm{p}}\pi r^2), \tag{4.14a}$$

where

$$q = 1 + \frac{p}{2(1+p)}\lambda_{\mathrm{p}}\pi\delta r. \tag{4.14b}$$

It satisfies all the standards for $r \geq 0$ meaning that it can also be applied to $\delta/2 \geq r \geq 0$. It is worth mentioning that $\lim_{\delta\to 0} q = 1$ and $\lim_{\delta\to 0} F(r)$ is the same as that of the HPPP with intensity λ_{p}, which is $1-\exp(-\lambda_{\mathrm{p}}\pi r^2)$. The proposed model fundamentally changes the exponent from a quadratic polynomial to a cubic polynomial, which better reflects the features of MHPP without losing the simplicity like $F(r)$ of a HPPP.

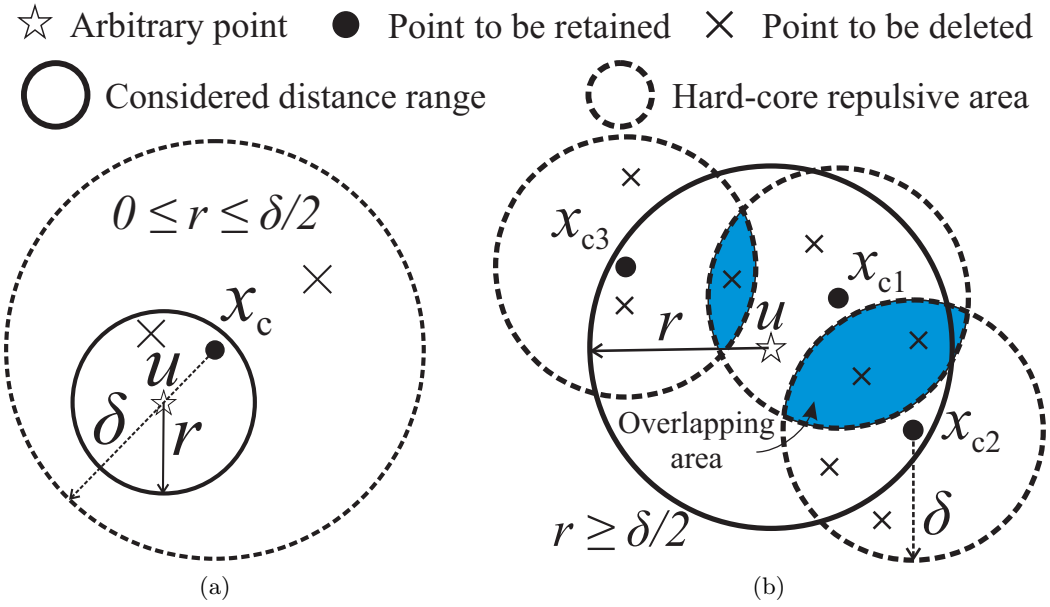


Figure 4.1: Examples for analysing the contact distance distribution of MHPP. (a) For $\delta/2 \geq r \geq 0$, a retained point deletes all the other primary points in the considered area; (b) For $r > \delta/2$, the probability of retaining multiple primary points depends on the overlapping hard-core areas of the to-be-retained points.

4.4.2 Nearest-neighbour Distance Distribution

The nearest-neighbour distance distribution $G(r)$ in terms of the cdf of a MHPP Φ_{m} at a target point $x_i \in \Phi_{\mathrm{m}}$ is formally defined as

$$G(r) = \mathbf{P}(\min_{x_j \in \Phi_{\mathrm{m}} \setminus x_i} \|x_i - x_j\| \leq r), \tag{4.15}$$

where $i \neq j$. Like $F(r)$, $G(r)$ does not depend on the location of x_i under consideration. No x_j can appear in $B(x_i, \delta)$ because of the existence of the hard-core repulsive area. Thus, different from that of a HPPP, $G(r)$ of a MHPP is equivalent to the probability of having at least one point in the annulus of inner radius δ and outer radius r as

$$G(r) = \mathbf{P}\Big(\Phi_{\mathrm{m}}(B(x_i, r) \setminus B(x_i, \delta)) \geq 1\Big). \tag{4.16}$$

The analysis of $G(r)$ is similar to that of $F(r)$ as explained before. It is also infeasible to derive a straightforward closed-form probability function considering retaining of multiple primary points with overlapping or non-overlapping hard-core areas in the annulus. For a pairwise distance $s = \|x_i - x_j\|$, the probability of retaining x_j conditioning on the event that

x_i is already retained, $h(s)$, can be calculated using $k(s)$. This can be formulated as

$$\mathbf{P}(x_j \in \Phi_{\mathrm{m}} \mid x_i \in \Phi_{\mathrm{p}}, x_j \in \Phi_{\mathrm{p}}, x_i \in \Phi_{\mathrm{m}}) = h(s) = \frac{k(s)}{p}. \tag{4.17}$$

$h(s)$ is larger than p for $2\delta > s > \delta$ (indicating the existence of an overlapping hard-core area) and converges to p at $s = 2\delta$, which implies no pairwise correlation between two points more than 2δ apart. In the example given in Fig. 4.2, the retaining probability of x_4 is $h(s)$ ($s = \|x_1 - x_4\|$) if considering the existence of the target point x_1. The conditional retaining probabilities of x_2 and x_3 are more complicated since they have multi-layer overlapping hard-core areas with each other and with the already retained x_1. Thus, when involving more to-be-retained points that can be contained in the annulus, it is impossible to derive a straightforward closed-form function.

However, like $F(r)$, it is also feasible to approximate $G(r)$ by a straightforward closed-form. The requirements of approximating $G(r)$ are similar to those of $F(r)$ except with $G(r) = 0$ for $[0, \delta]$. The proposed model is obtained by heuristic approximation satisfying the requirements and is shown as

$$G(r) = 1 - \exp\left(- qp\lambda_{\mathrm{p}}\pi(r^2 - \delta^2)\right), \tag{4.18a}$$

where

$$q = \begin{cases} 0, & 0 \le r \le \delta, \\ 1 + \frac{p}{3(1+p)}\lambda_{\mathrm{p}}\pi\delta r, & r > \delta. \end{cases} \tag{4.18b}$$

The piecewise function of q achieves the specification that the retaining probability of any primary point in $B(x_i, \delta)$ if $x_i \in \Phi_{\mathrm{m}}$ is always zero. $G(r)$ also approaches $1 - \exp(-\lambda_{\mathrm{p}}\pi r^2)$ as that of the parent HPPP when $\delta \to 0$.

4.4.3 *n*th Distance Distribution

Due to the thinning process, the number of surviving points in the MHPP Φ_{m} is smaller than (if $\delta > 0$) that of its parent HPPP Φ_{p}, resulting in Φ_{m} with intensity $p\lambda_{\mathrm{p}}$. Directly using the distance distribution models of Φ_{p}, and replacing the intensity by $p\lambda_{\mathrm{p}}$ neglects the existence of the hard-core area and lead to an inaccurate estimation of the distance distributions. However, the elegance of $F_n^{\mathrm{p}}(r)$ and $G_n^{\mathrm{p}}(r)$, the nth distance distributions of Φ_{p} with the same form as Equation (4.7), provides tractability in the mathematical analysis in its applications.

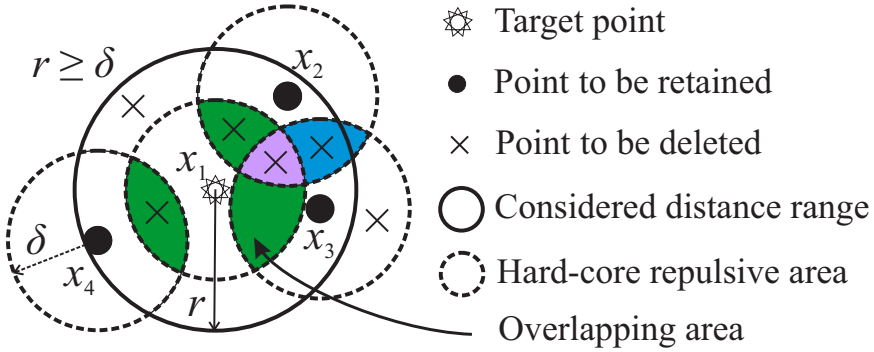


Figure 4.2: Example of analysing the nearest-neighbour distance distribution of MHPP. The existence of the target point raises the probability of retaining other primary points with overlapping hard-core areas. The to-be-retained points with overlapping hard-core areas also interact with each other, getting different retaining probabilities.

Thus, the approximation of nth ($n \geq 2$) distance distributions of Φ_{m} can be considered based on the distance distribution models of Φ_{p} and the differences of Φ_{m}.

Differences from a Parent HPPP

Considering $F_n(r)$, the probability of having at least n points in $B(u,r)$ increases along with the rise of r. However, there is an nth void space defined as $B(u,r_{\mathrm{v}}^F)$ (for an arbitrary u) which is the largest disc that cannot possibly contain n points of Φ_{m} (including those at the bound) due to the hard-core areas. It means that r in the interval $[0,r_{\mathrm{v}}^F]$ corresponds to $F_n(r) = 0$, where r_{v}^F is the radius of the nth void space. Compared with Φ_{m}, a HPPP with intensity $p\lambda_{\mathrm{p}}$ never has a zero probability at $r \neq 0$ for $F_n(r)$, which overestimates $F_n(r)$ in the interval $(0,r_{\mathrm{v}}^F]$ if the HPPP model is directly used without adjustment for the approximation.

To solve this issue, it is necessary to initially investigate r_{v}^F. The exploration of r_{v}^F is essentially to find the minimum radius of a container circle that can contain n points of Φ_{m} without having their pairwise distances less than δ. For $n \geq 2$ and $\delta > 0$, this falls into a complex packing problem, but can be well approximated by considering

$$r_{\mathrm{v}}^F = \min\left\{r \in (0,\infty) \,\middle|\, \sum_{k=0}^{\lceil \frac{r-\delta/2}{\delta} \rceil} \left\lceil \frac{\pi}{\arcsin(\frac{\delta/2}{r-k\delta})} \right\rceil + \mathbf{1}_{[0,\frac{\delta}{2})}(r - \delta\lfloor \frac{r}{\delta} \rfloor) \geq n \right\}, \tag{4.19}$$

where $\mathbf{1}_{[0,\frac{\delta}{2})}(\cdot)$ is an indicator function of $[0,\frac{\delta}{2})$ with the same definition given in Section 3.4.2. In Equation (4.19), the disc of considered radius r is split to an inner disc of the same centre and several concentric bordering annuli with the differences of the outer and inner radii being δ. The maximum number of points of Φ_{m} located at the outer circle of each annulus

is counted as the summation term in Equation (4.19). They are depicted as crosses located at the edges of the grey circles in the examples given in Fig. 4.3 (a), which shows the nth void space with r_{v}^F approximated by Equation (4.19). The maximum number of points at the outer circle (bound) of the residual inner disc after division is counted in the same way unless its radius is smaller than $\delta/2$, in which case the disc can contain one point of Φ_{m} as the second term in the set builder rule. An example of this type of point can be found in the fourth graph of Fig. 4.3 (a) and is depicted as the cross near the centre of the circle. The upper bound of the summation determines the number of the aforementioned annuli. r_{v}^F is determined to be r if the disc of radius r is just big enough to contain n points of Φ_{m}. For the special cases where $n = 1$ or $\delta = 0$, it is easily known that $r_{\mathrm{v}}^F = 0$.

It can be seen from Fig. 4.3 (a) that a minor error may exist where there are apertures that can contain points of Φ_{m} without affecting the arrangement by Equation (4.19), over-estimating r_{v}^F slightly. However, the minor error does not have a notable impact because Equation (4.19) needs to be further simplified to a closed form for the estimation of r_{v}^F such that it can be used for $F_n(r)$ approximation.

By $\arcsin(x) \approx x$, the difference of the number of points located at the outer circles of the adjacent annuli (or the inner disc) can be approximated to be 2π. This allows the maximum number of points of Φ_{m} that can be contained in the disc of radius r_{v}^F to be estimated by assuming an arithmetic progression. Thus, this gives the relationship as

$$n - 1 = \frac{r_{\mathrm{v}}^F}{2\delta}\Big(2\frac{\pi}{\delta/(2r_{\mathrm{v}}^F)} - 2\pi(\frac{r_{\mathrm{v}}^F}{\delta} - 1)\Big), \tag{4.20}$$

which derives

$$r_{\mathrm{v}}^F = \frac{-1 + \sqrt{1 + 4(n-1)/\pi}}{2}\delta. \tag{4.21}$$

On the left side of the estimation in Equation (4.20), -1 ensures $r_{\mathrm{v}}^F = 0$ when $n = 1$, indicating that the disc of radius r can contain at least one point of Φ_{m} for $r > 0$.

The feature of the nth void space is similar for $G_n(r)$, which is formally $B(x_i, r_{\mathrm{v}}^G)$ (for any $x_i \in \Phi_{\mathrm{m}}$) as the largest disc that cannot possibly contain n points of $\Phi_{\mathrm{m}} \setminus x_i$. Likewise, $G_n(r) = 0$ for $r \in [0, r_{\mathrm{v}}^G]$. Due to the hard-core area of x_i for $\delta > 0$, other points of Φ_{m} can not exist in $B(x_i, \delta)$ such that $r_{\mathrm{v}}^G \in [\delta, \infty)$ for any $n \geq 1$. Therefore, the problem is to find the minimum outer radius of a container annulus (or a circle for the equal inner and outer radii) that can contain n points of $\Phi_{\mathrm{m}} \setminus x_i$ without having their pairwise distances less than δ. Some examples are shown in Fig. 4.3 (b). For the cases where $n \leq 6$, $r_{\mathrm{v}}^G = \delta$ because any

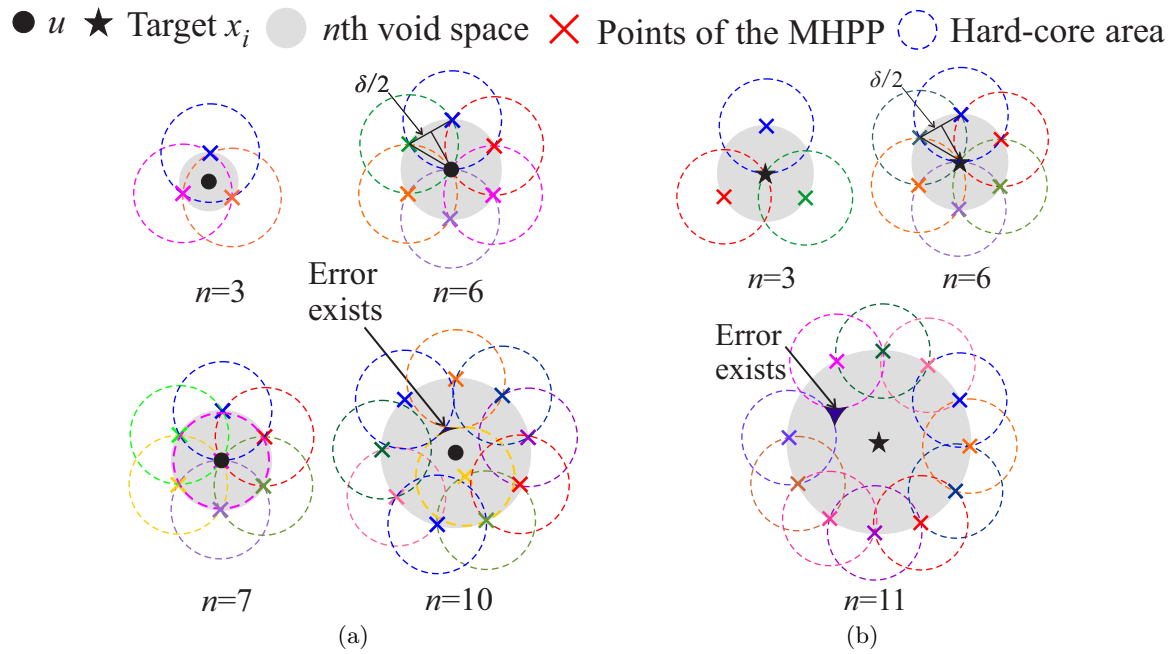


Figure 4.3: Examples of the nth void space containing n points of Φ_{m}. Special cases are shown with minor errors existing based on the approximation of Equation (4.19) or Equation (4.22), where one more point can be contained in either case. (a) For $F_n(r)$; (b) For $G_n(r)$.

$B(x_i, r_{\mathrm{v}}^G) \setminus B(x_i, \delta)$ for $r_{\mathrm{v}}^G \geq \delta$ can contain at least 6 points of $\Phi_{\mathrm{m}} \setminus x_i$. The packing problem for $n \geq 7$ and $\delta > 0$ can be solved by the approximation as

$$r_{\mathrm{v}}^G = \min \left\{ r \in [\delta, \infty) \,\middle|\, \sum_{k=0}^{\lceil \frac{r-\delta}{\delta} \rceil} \left\lceil \frac{\pi}{\arcsin(\frac{\delta/2}{r-k\delta})} \right\rceil \geq n \right\}. \tag{4.22}$$

Although there may also be minor errors existing as shown in the example, it does not have a significant impact as r_{v}^G should be further estimated by the same method to obtain a straightforward closed form for modelling $G_n(r)$. Assuming an arithmetic progression, the relation

$$n = \frac{r_{\mathrm{v}}^G}{2\delta} \left(\frac{\pi}{\delta/(2r_{\mathrm{v}}^G)} + 6 \right), \tag{4.23}$$

can deduce

$$r_{\mathrm{v}}^G = \frac{-3 + \sqrt{9 + 4\pi n}}{2\pi} \delta. \tag{4.24}$$

The final form of r_{v}^G also supports $r_{\mathrm{v}}^G = 0$ for $\delta = 0$.

Approximation of $F_n(r)$

Simplicity and tractability are the main reasons why the HPPP has been adopted widely for modelling wireless networks. In virtue of the straightforward closed-form $F_n^{\mathrm{p}}(r)$, its intensity

is replaced by the intensity of Φ_{m}, $p\lambda_{\mathrm{p}}$, after which $F_n^{\mathrm{p},p\lambda_{\mathrm{p}}}(r)$ is used as the basis for modelling $F_n(r)$ of Φ_{m}.

Compared with $\Phi_{\mathrm{p},p\lambda_{\mathrm{p}}}$ that $F_n^{\mathrm{p},p\lambda_{\mathrm{p}}}(r) > 0$ when $r > 0$, $F_n(r) > 0$ if $r > r_{\mathrm{v}}^F$. To realise the shift of the zero critical point, $F_n^{\mathrm{p},p\lambda_{\mathrm{p}}}(r)$ can be considered to shift to the positive direction for r_{v}^F at first. On reaching this, for a specific n, the shift is only a function of δ, a large value of which increases r_{v}^F. For a specific δ, this is particularly true when λ_{p} is large and $F_n^{\mathrm{p},p\lambda_{\mathrm{p}}}(r)$ diverges much from zero at a smaller r compared with $F_n(r)$. The circumstance is mainly caused by the increasing intensity $p\lambda_{\mathrm{p}}$ of $\Phi_{\mathrm{p},p\lambda_{\mathrm{p}}}$, which raises the probability of having at least n points, and there is no constraint of the maximum number of points that are possible to be contained in $B(u,r)$ for $\Phi_{\mathrm{p},p\lambda_{\mathrm{p}}}$. Thus, $F_n^{\mathrm{p},p\lambda_{\mathrm{p}}}(r)$ should be shifted for r_{v}^F to reflect the limitation of the hard-core areas. On the contrary, if λ_{p} is sufficiently small for a fixed δ, $F_n^{\mathrm{p},p\lambda_{\mathrm{p}}}(r)$ is of a longer quasi-zero region for the same r of the last case. $F_n(r)$ is constrained by the generation of the required number of points and the shift should be reduced to reflect the abated effect of r_{v}^F. Therefore, the shift should be dependent on r_{v}^F and correlated to λ_{p} and p (or δ^2). The linear correlation factor,

$$c = 1 - \exp(-5\lambda_{\mathrm{p}}\pi\delta^2), \tag{4.25}$$

is obtained by a heuristic approach and is demonstrated to be appropriate through simulation. The factor helps improve the accuracy of the model, but does not have as much impact as the previous parameters.

From the above analysis, considering a specific n, $F_n(r) \leq F_n^{\mathrm{p},p\lambda_{\mathrm{p}}}(r)$ for small r, where the number of possibly retained points in $B(u,r)$ is limited by the packing problem. Whereas for large r, the spacial limitation is relieved and primary points with pairwise distances less than 2δ but larger than δ have higher retaining probabilities than p, which indicates that more points of Φ_{m} are likely to be contained in $B(u,r)$ and $F_n(r) \geq F_n^{\mathrm{p},p\lambda_{\mathrm{p}}}(r)$. Hence, the monotonic increasing $F_n(r)$ versus r intersects with $F_n^{\mathrm{p},p\lambda_{\mathrm{p}}}(r)$ at $r = r_{\mathrm{e}}^F$, where the repulsive and absorbing effects of Φ_{m} are balanced in $F_n(r)$, behaving like $\Phi_{\mathrm{p},p\lambda_{\mathrm{p}}}$. The expected number of points of Φ_{m} in $B(u,r_{\mathrm{e}}^F)$, i.e. $p\lambda_{\mathrm{p}}\pi(r_{\mathrm{e}}^F)^2$, is discovered to be around $n-1$ at $r = r_{\mathrm{e}}^F$, which gives the relationship as

$$r_{\mathrm{e}}^F = \sqrt{\frac{n-1}{p\lambda_{\mathrm{p}}\pi}}. \tag{4.26}$$

After the shift of $F_n^{\mathrm{p},p\lambda_{\mathrm{p}}}(r)$ for modelling $F_n(r)$, it should be made equal to $F_n^{\mathrm{p},p\lambda_{\mathrm{p}}}(r)$ at

r_{e}^F. Then for $r > cr_{\mathrm{v}}^F$, $F_n(r)$ can be modelled as

$$F_n(r) = 1 - \exp\Big(- qp\lambda_{\mathrm{p}}\pi(r - cr_{\mathrm{v}}^F)^2\Big) \sum_{k=0}^{n-1} \frac{\Big(qp\lambda_{\mathrm{p}}\pi(r - cr_{\mathrm{v}}^F)^2\Big)^k}{k!}, \tag{4.27}$$

where to ensure $F_n(r_{\mathrm{e}}^F) = F_n^{\mathrm{p},p\lambda_{\mathrm{p}}}(r_{\mathrm{e}}^F)$,

$$q = \Big(\frac{r_{\mathrm{e}}^F}{r_{\mathrm{e}}^F - cr_{\mathrm{v}}^F}\Big)^2. \tag{4.28}$$

As aforementioned, $F_n(r) = 0$ for $r \in [0, cr_{\mathrm{v}}^F]$.

The proposed model aims at filling the gap of the lack of $F_n(r)$ for $n \geq 2$. It is worth mentioning that $\lim_{r \to cr_{\mathrm{v}}^F} F_n(r) = 0$ and $\lim_{r\to\infty} F_n(r) = 1$, making $F_n(r)$ continuous over $(0,\infty)$ and of cdf features. Additionally, $\lim_{\delta\to 0} F_n(r) = F_n^{\mathrm{p}}(r)$ meets the requirement that Φ_{m} approaches to the parent Φ_{p} without the thinning process as $\delta \to 0$. All the added parameters are independent of r, keeping $F_n(r)$ a straightforward, tractable but accurate closed form.

Approximation of $G_n(r)$

The approximation of $G_n(r)$ is similar to that of $F_n(r)$. Replacing the intensity of $G_n^{\mathrm{p}}(r)$ by $p\lambda_{\mathrm{p}}$, $G_n^{\mathrm{p},p\lambda_{\mathrm{p}}}(r)$ can be shifted to the positive direction for cr_{v}^G. As no points are in the hard-core area of a $x_i \in \Phi_{\mathrm{m}}$, r_{e}^G should be considered as the outer radius of an annulus with the inner radius of δ, in which the expected number of points of Φ_{m} is $n-1$. For $n \geq 2$. This derives

$$r_{\mathrm{e}}^G = \sqrt{\frac{n-1}{p\lambda_{\mathrm{p}}\pi} + \delta^2}, \tag{4.29}$$

$$q = \frac{(r_{\mathrm{e}}^G)^2 - \delta^2}{(r_{\mathrm{e}}^G - cr_{\mathrm{v}}^G)^2}. \tag{4.30}$$

Likewise, $G_n(r)$ for $r > cr_{\mathrm{v}}^G$ can be modelled as

$$G_n(r) = 1 - \exp\Big(- qp\lambda_{\mathrm{p}}\pi(r - cr_{\mathrm{v}}^G)^2\Big) \sum_{k=0}^{n-1} \frac{\Big(qp\lambda_{\mathrm{p}}\pi(r - cr_{\mathrm{v}}^G)^2\Big)^k}{k!}, \tag{4.31}$$

while for $r \in [0, cr_{\mathrm{v}}^G]$, $G_n(r) = 0$. All the requirements as aforementioned in modelling $F_n(r)$ are satisfied, giving an elegant closed-form $G_n(r)$ for $n \geq 2$.

4.4.4 Evaluation of the Distance Distributions

In this section, the virtues of the proposed models are demonstrated by comparing them with benchmark models. For the contact distance distribution, $F(r)$ of a HPPP with intensity $p\lambda_{\mathrm{p}}$ [80] is regarded as the benchmark. For the nearest-neighbour distance distribution, the method introduced in [33] is adopted. It fundamentally replaces the constant intensity λ_{p} of a HPPP with $\lambda_{\mathrm{p}}h(r)$, which converts it to an inhomogeneous PPP with a varying intensity being a function of the distance from the target point x_i. In this way, it only considers the condition that the to-be-retained points have overlapping hard-core areas with x_i but not with each other, neglecting most of the terms in the sum of the series. For the nth ($n \geq 2$) distance distributions of MHPP, the distance distribution functions of a HPPP with intensity λ_{p}, $F_n^{\mathrm{p},p\lambda_{\mathrm{p}}}(r)$ and $G_n^{\mathrm{p},p\lambda_{\mathrm{p}}}(r)$ are adopted as benchmarks for $F_n(r)$ and $G_n(r)$, respectively.

To quantify the accuracy and compare the models, Mean Squared Error (MSE) is adopted:

$$\mathrm{MSE} = \frac{1}{m}\sum_{i=1}^{m}(\hat{Y}_i - Y_i).^2, \tag{4.32}$$

where $\hat{Y}_i$ are the predictions produced by the models based on a set of input r_i as samples and Y_i are the observed results from the simulation based on the same input sample set. The number of samples denoted by m is selected to be 10000, which is big enough to generate smooth and stable curves. The samples of r_i are equally spaced and are taken from the main body of the functions that $F(r_i)$ or $G(r_i)$ is over the range of $[0.01, 0.99]$.

For the contact distance distribution and the nearest-neighbour distance distribution, since only the relative values of δ and λ_{p} matter, the models can be evaluated under different δ or λ_{p} with the other fixed. As can be seen from Fig. 4.4, the MSEs of the benchmark models both rise along with the increase of δ. This is because the increase makes the MHPP deviate more from its parent HPPP and the effect of neglecting the inter-point interaction by the benchmark becomes severe, raising the MSE consequently. Whereas the proposed models show almost zero MSEs comparatively and no variations for different r, significantly improving the accuracy. In the examples for $\lambda_{\mathrm{p}} = 0.02$ and $\delta = 5$ where p is a medium value, 0.5, as a representative, the benchmark models underestimate $F(r)$ or $G(r)$, revealing the sources of having larger MSEs. In addition to some accuracy improvement, the form of $G(r)$ is much simpler than the benchmark, providing a better tractability for mathematical analysis.

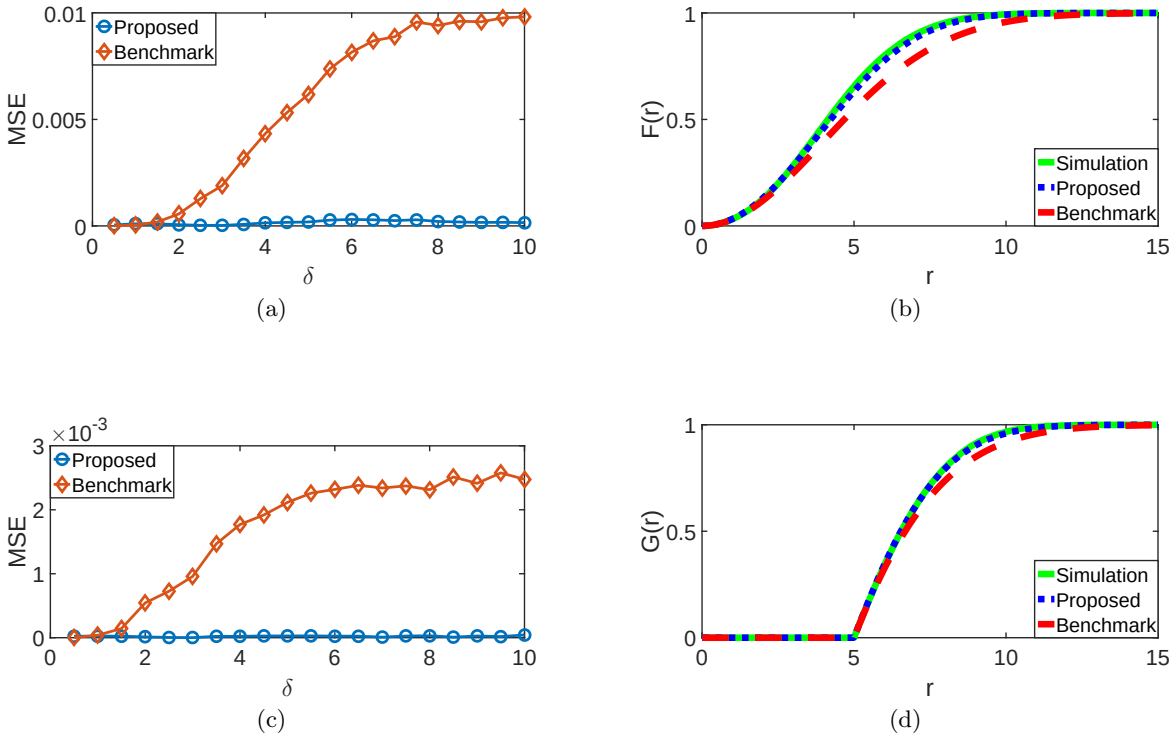


Figure 4.4: Comparisons and examples of the proposed models and the benchmark models for $F(r)$ and $G(r)$. (a) MSE comparison of the contact distance distribution $F(r)$ for $\lambda_p = 0.02$. (b) $F(r)$ examples, for $\lambda_p = 0.02$ and $\delta = 5$. (c) MSE comparison of the nearest-neighbour distance distribution $G(r)$ for $\lambda_p = 0.02$. (d) $G(r)$ examples, for $\lambda_p = 0.02$ and $\delta = 5$.

As can be seen from Fig. 4.5 (a) - (d), in the cases examined, the MSEs of the proposed models for both $F_n(r)$ and $G_n(r)$ are significantly reduced and are invariant versus n or δ compared with the benchmark HPPP models, which do not take the thinning process into consideration. This is especially obvious for MSE versus δ, the increase of which deviates the HPPP models from the observed results because of the thinning process with larger hard-core distances. As λ_p is a relative value to δ, it can be fixed when investigating the effect of varying δ as presented in Fig. 4.5 (c) and (d).

The applications of the distance distribution functions usually involve the derivative of a cdf, i.e. the probability density function (pdf). The derivatives of $F_n(r)$ and $G_n(r)$ are denoted as $f_n(r)$ or $g_n(r)$, respectively, and are shown in Fig. 4.5 (e) and (f) as the examples. As can be seen, the proposed models fit the simulation results very well since $f_n(r)$ and $g_n(r)$ are more concentrated at medium r. This validates the aforementioned analysis that the existence of hard-core areas repulses at least n points to be contained if r is too small, which refers to a packing problem. However, it also elevates the retaining probabilities of primary points with overlapping hard-core areas and with pairwise distances larger than δ.

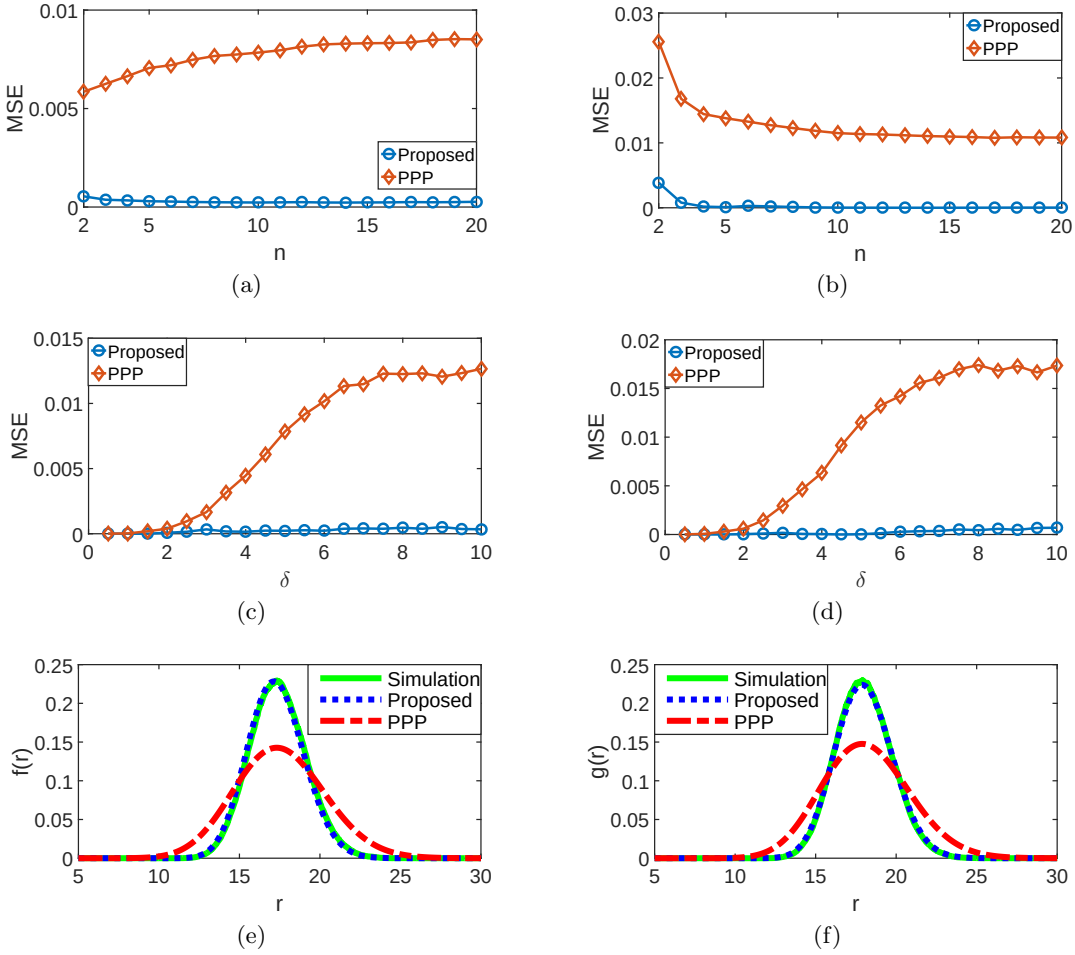


Figure 4.5: Comparisons and examples of the proposed models and the HPPP model for the nth distance distribution. (a) MSE versus n, for $F_n(r)$, $\lambda_\mathrm{p} = 0.02$ and $\delta = 5$. (b) MSE versus n, for $G_n(r)$, $\lambda_\mathrm{p} = 0.02$ and $\delta = 5$. (c) MSE versus δ, for $F_n(r)$, $\lambda_\mathrm{p} = 0.02$ and $n = 10$. (d) MSE versus δ, for $G_n(r)$, $\lambda_\mathrm{p} = 0.02$ and $n = 10$. (e) $f_n(r)$ examples, for $\lambda_\mathrm{p} = 0.02$, $\delta = 5$ and $n = 10$. (f) $g_n(r)$ examples, for $\lambda_\mathrm{p} = 0.02$, $\delta = 5$ and $n = 10$.

This property increases the probability of having at least n points for medium r and lower the probability for small and large r that is predicted by the HPPP model. These all together exhibit a more concentrated pdf compared with the HPPP models of the same intensities, $p\lambda_\mathrm{p}$.

4.5 Applications: a Case Study

The applications of the distance distributions comprise any analysis where the distances between transmitters and receivers matter. The conventional directions include but not limit to the analysis of interference, outage probability (or coverage probability), routing [81], transmission power, localisation, cell planning, connectivity, etc. This section analyses the

distribution of transmission power under open-loop power control, which directly adopts the derivative of $F(r)$, $f(r)$.

Assuming Rayleigh fading [27, 82], the received power at a constant distance r from a transmitter can be modelled as $P_t h r^{-\alpha}$, where P_t is the transmitted power, h is the random chain gain exponentially distributed with mean $1/\mu$ and α is the path loss exponent. Then for open-loop power control achieving a signal-to-noise ratio of β with noise denoted as σ^2, the cdf of P_t is

$$\begin{aligned}\mathbf{P}(P_t \leq \theta | r) &= \mathbf{P}\Big(\frac{\beta\sigma^2}{h r^{-\alpha}} \leq \theta | r\Big) = 1 - \mathbf{P}\Big(h \leq \frac{\beta\sigma^2 r^\alpha}{\theta} | r\Big) \\ &= \exp(-\frac{\mu\beta\sigma^2 r^\alpha}{\theta}).\end{aligned} \tag{4.33}$$

$f(r)$ is used in relaxing r as

$$\mathbf{P}(P_t \leq \theta) = \int_0^\infty \exp(-\frac{\mu\beta\sigma^2 r^\alpha}{\theta}) f(r) \mathrm{d}r. \tag{4.34}$$

The cdf of P_t is compared between the MHPP and the HPPP with the same intensity for different α in Fig. 4.6. The results indicate less transmission power in networks modelled by a MHPP due to its more regular pattern. The difference is significant especially when the path loss is large for the same r with a higher α, which also demonstrates the impact of inaccuracy in the system analysis if using the PPP based approximation to capture the distance distributions in networks model by MHPPs. The investigation of transmission power is particularly important at the user side due to the limited battery lives of mobile devices. It relates to the topic of the thesis and will be given further analysis in Chapter 7.

4.6 Conclusion

In this chapter, the distance distributions of HPPP and the properties of type II MHPP are first studied. Based on the insights, the rigorous derivation of the contact distance distribution $F(r)$ for type II MHPP when $0 \leq r \leq \delta/2$ is proposed. The infeasibility of deriving complete closed-form distance distributions is analysed. Two straightforward models are presented for the contact distance distribution $F(r)$ and the nearest-neighbour distance distribution $G(r)$. The nth ($n \geq 2$) distance distributions, $F_n(r)$ and $G_n(r)$, are also analysed and approximated with tractable closed-forms. Some of the results will be directly used in Chapter 5 and Chapter 7.

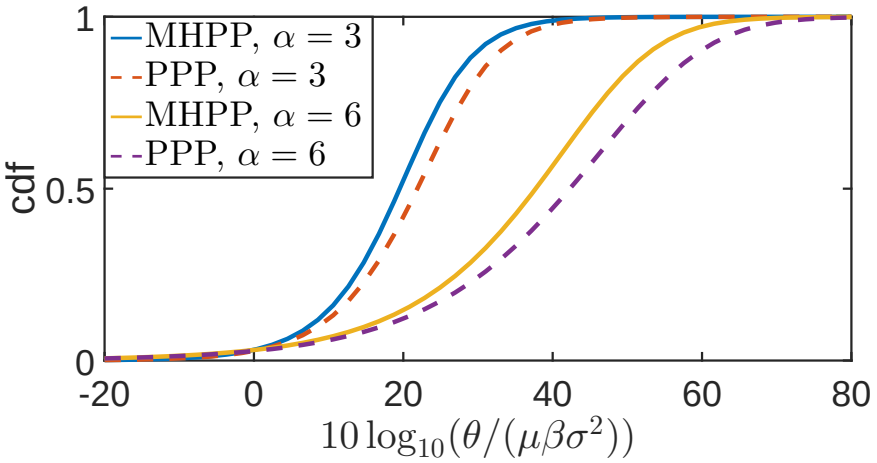


Figure 4.6: Comparison of the distributions of the controlled transmission power between networks modelled by the MHPP and the HPPP with the same intensity.

The MSEs of the proposed models are nearly zero, demonstrated to be significantly better than the benchmark models in terms of accuracy. The examples of either the cdf form or the pdf form of the distance distributions are provided. With guaranteed accuracy, the pdf form of the proposed models can be effectively used for the further mathematical analysis of cellular networks modelled by MHPPs, such as transmission power consumption studied as an example. Moreover, the applications of the proposed models are not only limited to the field of wireless communication, but can also be applied in any fields where MHPP is used and distance-related metrics are critical.

Chapter 5

Adaptive Traffic Perception Based Sleep Mode Operation

Contents

5.1 Introduction

As reviewed in Chapter 2, sleep mode operation is a vital part of on-demand radio resource provisioning. It is expected to reduce the overall network power consumption by making a SCN switch its BSs to low-power sleep modes when the network load is low. For a SCN BS with the capability for frequent switching, fast reaction sleep mode operation can be applied to realise a timely response to varying traffic volume for the maximisation of energy saving. As mentioned in the hypothesis of the thesis, appropriate traffic information is essential in sleep mode operation. In a SCN with a distributed architecture like LTE, the challenge

of designing an appropriate sleep mode operation scheme is to enable a BS to identify the network and traffic conditions without having global knowledge. In other words, a BS should be aware of the traffic it should support and the traffic afforded by other active BSs to determine its state and the opportune moment for state transition. Due to the distributed architecture, the inter-BS connectivity is usually limited to the local area of any BS, where effective traffic information should be collected for the aforementioned reasoning.

This chapter proposes a novel sleep mode operation scheme based on Adaptive Traffic Perception (ATP) for on-demand radio resource provisioning. The concept of ATP and its implementation in a practical SCN are introduced in Section 5.2. The sleep mode operation scheme proposed takes it as the basis and is presented in Section 5.3. The performance of the scheme is evaluated in Section 5.4. Furthermore, the scheme is also analysed with the MHPP model, which is a more realistic stochastic model to capture the spatial distribution of SCN BSs. The analysis and the comparison are in Section 5.5, which directly use a result in Chapter 4. Finally, the conclusion is drawn in Section 5.6. Owning to the distributed architecture under discussion, the LTE notations are adopted, where an eNB refers to a BS and a UE denotes a user.

5.2 Adaptive Traffic Peception

To overcome the drawback of a distributed architecture, an eNB, as a decision-making agent, can understand the system status by considering the conditions of its neighbour eNBs. Specifically in sleep mode operation, an eNB can measure the traffic volume of its local area, to which it is responsible to provide wireless access. The measured traffic volume is an appropriate decision-making quantity for determining the state of an eNB. However, the local area of an eNB in a SCN with a frequently changed topology due to sleep mode operation is usually ambiguous. This problem is solved in this section and the mechanism of Adaptive Traffic Perception (ATP) is described.

5.2.1 Concept

In ATP, each eNB is designed with a circular monitoring area, which is centred at the eNB itself with a particular monitoring radius. For a specific eNB, its monitoring radius takes the value of the distance from the eNB to its nearest active eNB. Each eNB monitors and counts the active data transmission links within its own monitoring area as the traffic volume in

its local area. The concept of ATP is straightforward, but clearly defines the local area of an eNB, which adaptively changes according to the states of its adjacent eNBs. Intuitively, the monitoring radius of an eNB is large when its nearest active eNB is far away, meaning that there are no nearby active eNBs to support the traffic and a wider local area should be defined to monitor the traffic. Conversely, the monitoring radius of a target eNB is relatively small if there is an active eNB nearby indicating that it provides wireless services and the target eNB can be focused on a smaller local area. As an eNB is always only responsible for the traffic in its local area, it is the definition of the local area which matters. It is not necessary for an eNB to consider the traffic in a distant area and the eNB operational states there. With an adaptive monitoring area, an eNB can gauge the local traffic volume as the decision-making quantity in the sleep mode operation.

5.2.2 Practical Implementation

Being straightforward in concept, its application in the real system is more complex. Practically, the determination of the monitoring radius can be done through power measurement because power attenuation is mainly caused by path loss as a function of distance, which can be represented as a distance dependent function, $l(d)$ (in decibel). In the ultra-dense SCN considered, the eNBs are denoted by a set $\mathcal{B} = \{B_i\} = \mathcal{B}^{\text{on}} \cup \mathcal{B}^{\text{off}}$ ($b \in \mathbb{N}^*$), where $\mathcal{B}^{\text{on}}$ and $\mathcal{B}^{\text{off}}$ are active and sleeping eNBs, respectively. The distance between B_i and B_j is denoted as $D_{i,j}$ ($i,\ j \in \mathbb{N}^*,\ j \neq i$). For an active eNB B_j, it periodically broadcasts an indicator signal with the pre-defined power P_{ref} to inform the other eNBs of its active state. Each eNB B_i receiving this message independently measures and records power of the received signal $P_{i,j}^{(\text{B})}$ in the memory, which realises recording $D_{i,j}$ because of the relationship as

$$P_{i,j}^{(\text{B})} = P_{\text{ref}} - l(D_{i,j}), \tag{5.1}$$

where $P_{i,j}^{(\text{B})}$ and P_{ref} are in dBm corresponding to $l(D_{i,j})$ in dB.

Consequently, distance is stored in form of power, which is easy to be realised in the practical system. The monitoring area of B_i can thus be depicted by the monitoring reference power R_i, corresponding to its equivalent monitoring radius M_i. The calculation of R_i is periodically processed by B_i as

$$R_i = \min\Big(\max_j(P_{i,j}^{(\text{B})}, R_{\min}), R_{\max}\Big), \tag{5.2}$$

where $R_{\min}$ is specifically defined to model the limited ability of an eNB that it cannot confidently sense all the signals of arbitrarily low power. $\max_j(P_{i,j}^{(\mathrm{B})}, R_{\min})$ forces any received signal with power $P_{i,j}^{(\mathrm{B})}$ lower than $R_{\min}$ to be filtered out. $R_{\min}$ can be estimated following $R_{\min} = P_{\mathrm{ref}} - l(M_{\max})$ for the comparison with $P_{i,j}^{(\mathrm{B})}$, where $M_{\max}$ indicates the maximum equivalent monitoring radius. With the knowledge of the maximum coverage area of a SCN eNB treated as $M_{\max}$, $R_{\min}$ can be derived. On the other hand, $R_{\max}$ in form of power is to set a minimum equivalent monitoring radius $M_{\min}$ following $R_{\max} = P_{\mathrm{ref}} - l(M_{\min})$. Due to the variations of the radio environment, there may be a path between B_i and B_j with extremely low path loss, yielding much higher $P_{i,j}^{(\mathrm{B})}$ compared with other paths, most of which have normal propagation conditions, such as random shadowing. Bounding the upper end of R_i by $R_{\max}$ avoids this situation where the equivalent monitoring radius is reduced to an arbitrary small value. If B_i does not receive the indicator signal from B_j for a period of time, B_i removes the corresponding record $P_{i,j}^{(\mathrm{B})}$ from its memory, after which B_i repeats the above computation to update R_i.

At the UE side, a UE $u_m \in \mathcal{U}$ ($m \in \mathbb{N}^*$) being delivered services periodically broadcasts a reference signal for each resource allocation unit, which signifies an active link $u_{m,a}$ ($a \in \mathbb{N}^*$). Each eNB B_i records $u_{m,a}$ in its memory $\mathcal{S}_i$ if B_i can sense it, that is, power (in dBm) of the received reference signal at B_i sent by u_m for link a, $P_{i,m,a}^{(\mathrm{U})}$, is not less than the pre-defined $R_{\min}$, which corresponds to the maximum equivalent monitoring radius $M_{\max}$. This process can be formulated as

$$\mathcal{S}_i = \{u_{m,a} : P_{i,m,a}^{(\mathrm{U})} \geq R_{\min}\}. \tag{5.3}$$

If the transmitting power of the reference signal for the link a sent by u_m is the same as the power of the indicator signal sent by each eNB, P_{ref}, $P_{i,m,a}^{(\mathrm{U})}$ can be calculated as

$$P_{i,m,a}^{(\mathrm{U})} = P_{\mathrm{ref}} - l(d_{i,m}) \tag{5.4}$$

where $d_{i,m}$ denotes the distance between u_m and B_i. Based on the prerequisite, each eNB B_i can thus determine whether u_m is within its monitoring area by comparing $d_{i,m}$ and M_i, which is to compare $P_{i,m,a}^{(\mathrm{U})}$ with R_i (of power form) practically. B_i moves $u_{m,a}$ into its monitoring list $\mathcal{W}_i$ if the receive power is greater than or equal to its R_i as

$$\mathcal{W}_i = \{u_{m,a} : P_{i,m,a}^{(\mathrm{U})} \geq R_i,\ u_{m,a} \in \mathcal{S}_i\} \tag{5.5}$$

Conversely, if u_m departs when finishing data transmission over link a, B_i removes $u_{m,a}$ from

its memory $\mathcal{S}_i$ and monitoring list $\mathcal{W}_i$ if $u_{m,a}$ exists. By creating $\mathcal{S}_i$, B_i can re-investigate whether $P^{(\mathrm{U})}_{i,m,a}$ of any detectable UE is greater than or equal to R_i to update the total number of active links within its monitoring area whenever R_i is changed. The computations of R_i and $\mathcal{W}_i$ are both periodical to ensure that B_i can update the monitoring as soon as there is a state transition nearby.

Therefore, the state transformations of eNBs in real time lead to adaptive adjustments of their monitoring area in the form of monitoring reference power, based on which the ranges of traffic perception are adjusted accordingly and the number of active links counted by all eNBs are different. At any time, the traffic perceived by each eNB is limited to its local area defined by the states of other near eNBs, which achieves the objective of considering the conditions of the system while meeting the local demand. The configuration of ATP does not bring much adjustment to the current cellular system. Similar forms of the indicator signals and the reference signals can all be found in the current control signals. P_{ref} can be set to power of these control signals. $R_{\min}$ can be configured with the maximum coverage area of a SCN eNB as the reference. $R_{\max}$ can be chosen to be a similar value to the ISD of the SCN.

5.3 Self-driven Sleep Mode Operation

The sleep mode operation of each eNB based on ATP is self-driven without any extra information exchanged among eNBs. An eNB B_i activates itself if its perceived traffic, which is the total number of active links within its monitoring area, $|\mathcal{W}_i|$, is continuously above a pre-defined threshold, W_{th} for a specified duration, H.

On the other hand, it will be prepared to be switched to the sleep mode if $|\mathcal{W}_i|$ continuously keeps below or equal to W_{th} for the duration H and the load of B_i in terms of the number of occupied resource allocation units, L_i, also keeps below or equal to W_{th} continuously for the duration H. A UE may connect out of the monitoring area of B_i. It may also be within the monitoring area of B_i, but connecting to another eNB. These two cases make L_i differ from $|\mathcal{W}_i|$. L_i is taken into consideration in the condition of switching to sleep modes for an extreme case of low offered traffic levels, where L_i is above the threshold, but with the UEs connecting out of the monitoring area of B_i because fewer eNBs are active. In this case, there may be small $|\mathcal{W}_i|$ below the threshold, but B_i should not be switched to the sleep mode because of its high load. For high traffic levels, L_i is usually lower than $|\mathcal{W}_i|$ because the monitoring area of B_i is usually larger than its Voronoi cell, making more monitored active

links than occupied radio resource units. Hence, L_i satisfies the requirement of switching to the sleep mode if $|\mathcal{W}_i|$ is below the threshold as well. In this case, decision-making of both activation and deactivation mostly relies on comparing $|\mathcal{W}_i|$ with the threshold, reducing the ping-pong effect, where an eNB meets the condition of activation instantly after switching to the sleep mode or vice versa. B_i sends a request to its nearest active eNB B_j after it is prepared to be switched to the sleep mode. B_j sends back an agreement message if B_j is not prepared to be switched to the sleep mode. B_i only turns to the transient state for switching to the sleep mode only after it receives the agreement from B_j. This 'handshake' process is to prevent concurrent switching of multiple nearby eNBs, which is a hazard in a distributed system with all eNBs making decisions independently. In an extreme case where there is only one active eNB in a SCN, it has no active nearest eNB to send back an agreement for switching to the sleep mode. Limited by the 'handshake' process, it always keeps active to ensure connectivity in the area covered by the SCN if the other eNBs are hard to be activated due to extreme low traffic levels in their monitoring areas. Before an activation or a 'handshake', an eNB has to clear the time counter, which continuously counts the time to H for a condition.

After state transitions, an eNB B_i may find R_i changed due to the switching on/off of a near eNB, which alters the distance to the nearest eNB of B_i. $\mathcal{W}_i$ is consequently adjusted as long as R_i changes. Fig. 5.1 shows an example of the ATP based sleep mode operation, assuming that the equivalent monitoring radius M_i estimated based on the path loss function is equal to the real distance and $W_{\text{th}} = 1$. The left part of the figure presents the traffic perceptions made by B_1, B_2 and B_3. M_1 and M_2 depend on the distances to the nearest active eNBs B_3 and B_4 of B_1 and B_2, respectively, which yield $\mathcal{W}_1 = \{u_{3,1}, u_{4,1}, u_{5,1}\}$ and $\mathcal{W}_2 = \{u_{2,1}, u_{5,1}\}$ with $|\mathcal{W}_1| = 3$ and $|\mathcal{W}_2| = 2$. M_3 takes the value of $M_{\max}$ because B_3 is far away from its nearest active eNB B_4, which results in $\mathcal{W}_3 = \{u_{2,1}\}$ with $|\mathcal{W}_3| = 1$. As $|\mathcal{W}_1| > 1$, $|\mathcal{W}_2| > 1$ and $|\mathcal{W}_3| = L_3 = 1$, B_1, B_2 activate themselves and B_3 switches itself to the sleep mode after the conditions are continuously met for H. After state transitions, the equivalent monitoring radii are adaptively adjusted as shown in the right part of the figure, where M_1 and M_3 are reduced due to the activations of B_2 and B_1, respectively. The adaptations reconfigure the monitoring lists as $\mathcal{W}_1 = \{u_{3,1}, u_{4,1}\}$ and $\mathcal{W}_3 = \varnothing$, reflecting the changes of the local areas. The overload of B_4 is mitigated.

Concretely, the ATP based sleep mode operation algorithm enables each eNB to adaptively define its own local area, where the traffic level is perceived by counting the active links. The direct traffic information exploited by the mechanism are the counted number fo

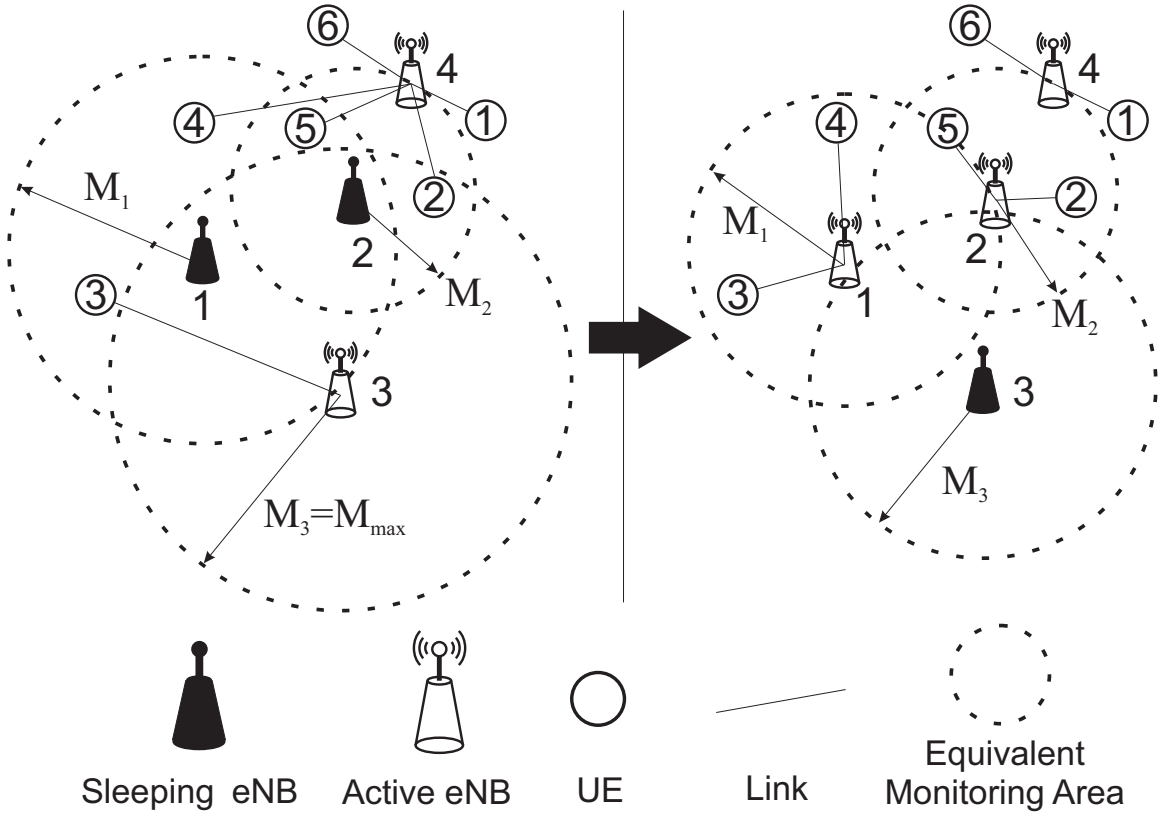


Figure 5.1: Example of the ATP based sleep mode operation. The left part shows the traffic perception before state transitions, which are made based on the conditions of the ATP based sleep mode operation. The right part shows the adaptive adjustment after the state transitions.

active links and the loads. The distance to the nearest active eNB is also indirectly used in the format of the received power, indicating the traffic levels around the local area of an eNB. For a specific eNB B_i, if the distance to its nearest active eNB becomes smaller as a sleeping eNB is activated, B_i regards it as a situation where part of its previous local area has been supported by the newly activated eNB and therefore will reduce its monitoring radius. After shrinking the monitoring area, the traffic level perceived by B_i is reduced and B_i is easier to be deactivated or harder to be activated, matching the fact that more eNBs nearby are active and B_i is responsible for a smaller local area. Thus, from a global view, the adaptation of traffic perception yields different requirements in traffic density for an eNB to take action, either activation or deactivation. More active eNBs in the network with shrinking monitoring areas need higher traffic densities to activate another eNB with a fixed W_{th}. By this mechanism, eNBs can make decisions depending on the local traffic density and the partial system conditions even with a distributed architecture. The pseudo code of the ATP based sleep mode operation is presented as follows.

for a $B_i \in \mathcal{B}$, B_i **do**
  if indicator signal periodically received from B_j **then**
    Measures and updates $P_{i,j}^{(\mathrm{B})}$
  else
    $P_{i,j}^{(\mathrm{B})} = 0$
  end if
  Computes and updates $R_i = \min\left(\max_j(P_{i,j}^{(\mathrm{B})}, R_{\min}), R_{\max}\right)$
  Computes and updates $\mathcal{S}_i = \{u_{m,a} : P_{i,m,a}^{(\mathrm{U})} \geq R_{\min}\}$
  Computes and updates $\mathcal{W}_i = \{u_{m,a} : P_{i,m,a}^{(\mathrm{U})} \geq R_i, u_{m,a} \in \mathcal{S}_i\}$
end for

for a $B_i \in \mathcal{B}^{\mathrm{on}}$, B_i **do**
  Periodically broadcasts an indicator signal with power P_{ref}
  if $|\mathcal{W}_i| \leq \mathcal{W}_{\mathrm{th}}$ continuously for $H \wedge L_i \leq \mathcal{W}_{\mathrm{th}}$ continuously for H **then**
    Prepares to switch to the sleep mode
    Clears the time counters for counting to H
    Sends a request for switching to the sleep mode to $B_j = \arg\max_{B_j \in \mathcal{B}^{\mathrm{on}},\ j \neq i} P_{i,j}^{(\mathrm{B})}$
    if an agreement message received from B_j **then**
      Switches to the sleep mode
    else
      Ceases the preparation for switching
    end if
  end if
  if a request for switching to the sleep mode received from B_k $(k \neq i)$ **then**
    if not in the preparation for switching **then**
      Sends back an agreement message to B_k
    end if
  end if
end for

for a $B_i \in \mathcal{B}^{\mathrm{off}}$, B_i **do**
  if $|\mathcal{W}_i| > \mathcal{W}_{\mathrm{th}}$ continuously for H **then**
    Clears the time counter for counting to H
    Activates B_i

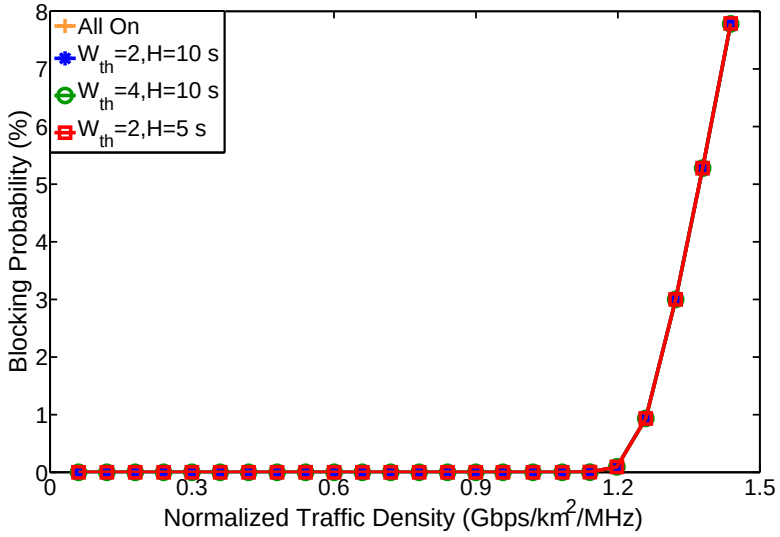
35: **end if**
36: **end for**
37:
38: **for** a $u_m \in \mathcal{U}$, u_m **do**
39: **if** transmitting or receiving by an active link a **then**
40: Periodically broadcasts a reference signal with power P_{ref} for the link a
41: **end if**
42: **end for**

5.4 Performance Evaluation

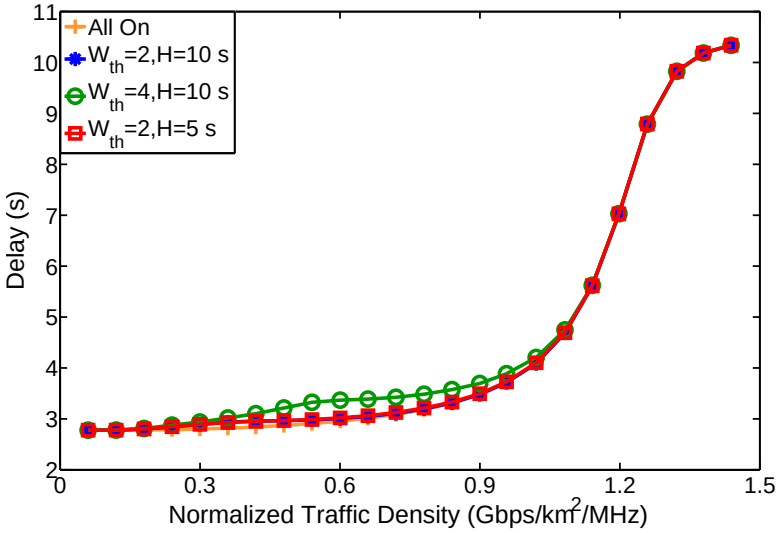
The performance of the proposed ATP based sleep mode operation is evaluated using the models introduced in Chapter 3. The SCN is modelled with ideal spatial distributions given in Section 3.2.1, where 6 rings of hexagonal cells with 91 eNBs construct the system and the ISD is chosen to be 10.7 metres. The eNB power consumption model in Section 3.2.7 is adopted for the evaluation of the network power consumption. The rest of the system is modelled as presented in Section 3.2. 100 systems are built for simulation using the simulation techniques presented in Section 3.3.

5.4.1 Trade-off between QoS and Power Consumption

In Fig. 5.2, the QoS in terms of blocking probability and delay for the proposed ATP based sleep mode operation algorithm is displayed. The shown results are based on controlled variables, i.e. fixing one parameter and varying the other. As can be seen from Fig. 5.2 (a), all the configurations of the algorithm yield the same average blocking probability as the system without sleep mode operation. As a threshold for activation and deactivation, W_{th} controls the average number of the active eNBs in a SCN, which directly affects the number of radio resources in the SCN. With a lower W_{th} and more high quality radio resources, the small cell system delivers services with higher transmission rates and achieves the same delay as the system without sleep modes as shown in Fig. 5.2 (b). H is a parameter to control the frequency of state transitions and is also prevents the eNBs from overreacting to temporary traffic variations. With a unique H for both activation and deactivation, the requirements of continuously keeping above or below the switching threshold are balanced for switching on and switching off, which therefore has a limited effect in QoS. The reduction in H as shown



(a)



(b)

Figure 5.2: QoS for the ATP based sleep mode operation against normalised traffic density. (a) Blocking probability. (b) Delay.

in Fig. 5.2 does not deviate QoS from the one with the same W_{th}. The other combinations of the values of the parameters are not shown because the existing ones have already represented the problems considered.

The trade-off can be observed by contrasting QoS with area power consumption presented in Fig. 5.3. The lower bound of area power consumption is represented as a dashed line parallel to the x-axis at the bottom of the figure. It is calculated assuming that all the eNBs are switched to sleep modes and indicates the maximum reduction in area power consumption, which can never be achieved, but signifies the upper bound of the algorithm performance. Conversely, the system without sleep mode operation is at full power and pro-

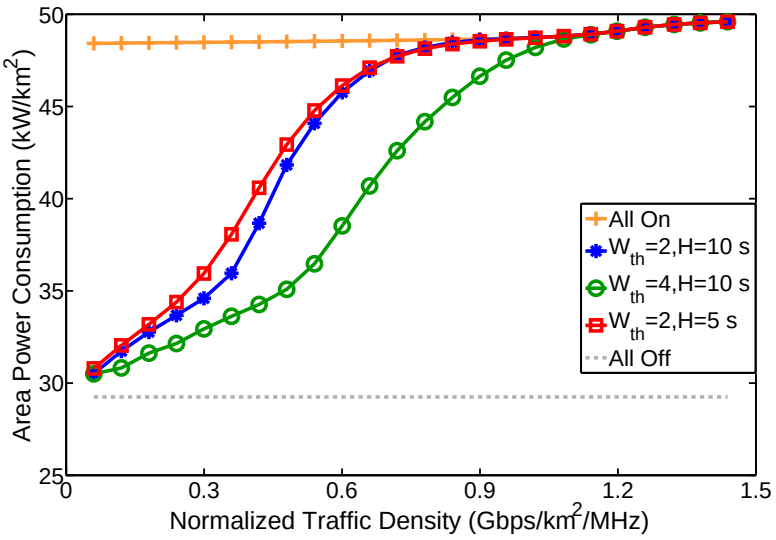


Figure 5.3: Area power consumption for the ATP based sleep mode operation against normalised traffic density.

vides the maximum area power consumption of a SCN, which is positively correlated with the average traffic density. It is the most obvious at high traffic levels, where an extreme high average delay and the corresponding high loads exist. At these traffic levels, the ATP based sleep mode operation algorithm is able to switch all the eNBs on to provide enough radio resources, showing a convergence to area power consumption of a system without sleep mode operation. At a low traffic level, the ATP based sleep mode operation algorithm achieves a reduction in area power consumption by around 37% compared with the system without sleep mode operation. Notably, although H only has a minor effect on QoS, its reduction slightly raises area power consumption by a higher switching frequency, which contributes to higher energy consumption during transient states estimated by assuming a transient state of medium energy consumption introduced in Section 3.2.7. W_{th} controls the average number of the active eNBs in a SCN as aforementioned, thus determines area power consumption. Its increase reduces area power consumption because of fewer active eNBs in the SCN. However, it affects the system performance with higher delay and generates the main trade-off of a sleep mode operation algorithm. More discussions of W_{th} is provided in Section 5.5.2.

5.4.2 Impact of the ATP Implementation

Although a reasonable ATP implementation has been suggested, the effects of extreme cases are investigated as shown in Fig. 5.4, where different extreme $M_{\max}$ or $M_{\min}$ are chosen as representatives. $R_{\max}$ or $M_{\min}$ exists to prevent an eNB from having a shrunk monitoring

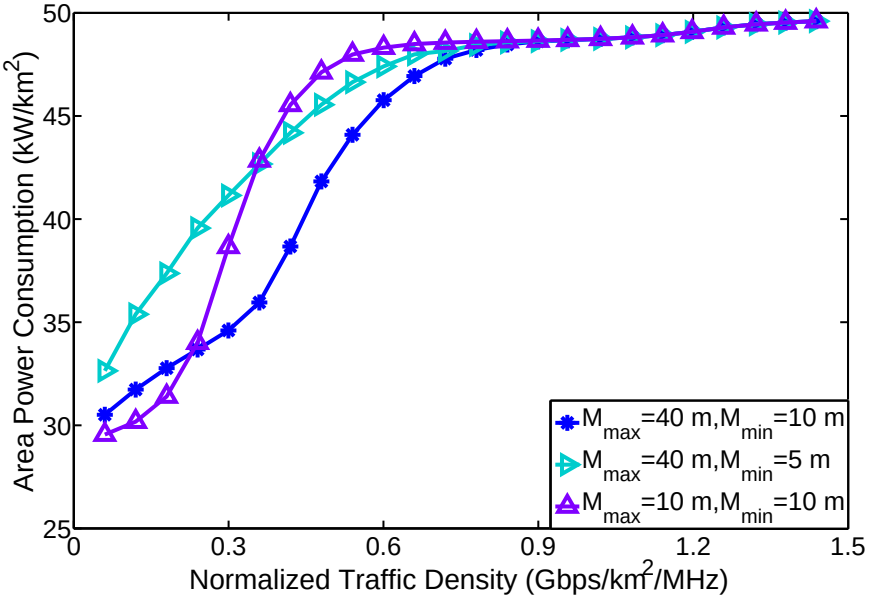


Figure 5.4: Area power consumption for the different configurations of the equivalent monitoring radius against normalised traffic density.

area caused by extreme low path loss between itself and its nearest active eNB. It is more likely to occur at high traffic levels with small $\overline{M_i}$ or large $\overline{R_i}$, which leads to fewer active eNBs in a SCN and corrupts QoS. To compensate the impact if $M_{\min}$ is set to a low value such as 5 metres in the figure, W_{th} has to be adjusted to maintain QoS for high traffic levels, but thereafter raises area power consumption. $R_{\min}$ or $M_{\max}$ are to model the sensing ability of an eNB that it cannot sense any arbitrarily low power. Large $R_{\min}$ or small $M_{\max}$ indicates the inability of having a large monitoring area required at low traffic levels. The example with $M_{\max}$ equal to 10 metres is an extreme case where the monitoring radius is fixed to a low end. It limits the maximum number of active links that can be included in a monitoring area and yields fewer active eNBs at low traffic levels. W_{th} also needs to be adjusted to maintain QoS, which brings about higher area power consumption at high traffic levels instead.

5.5 On Non-ideal Topology

Last section presents the performance evaluation of the ATP based sleep mode operation on an ideal network layout consisting of hexagonal grids. It is suitable to model the SCNs in special cases, where the positions of eNBs are carefully planned. In more general cases, stochastic models are more appropriate to capture the randomness in eNB placement of the SCNs as reviewed in Chapter 2. This section compares the system performance with or without ATP based sleep mode operation on the hexagonal grid model and the MHPP model. The proposed

algorithm is also analysed on a SCN modelled with the MHPP, where the relationship among QoS, network power consumption, offered traffic and a algorithm parameter is captured. The analysis involves a direct application of $g_n(r)$ proposed in Chapter 4.

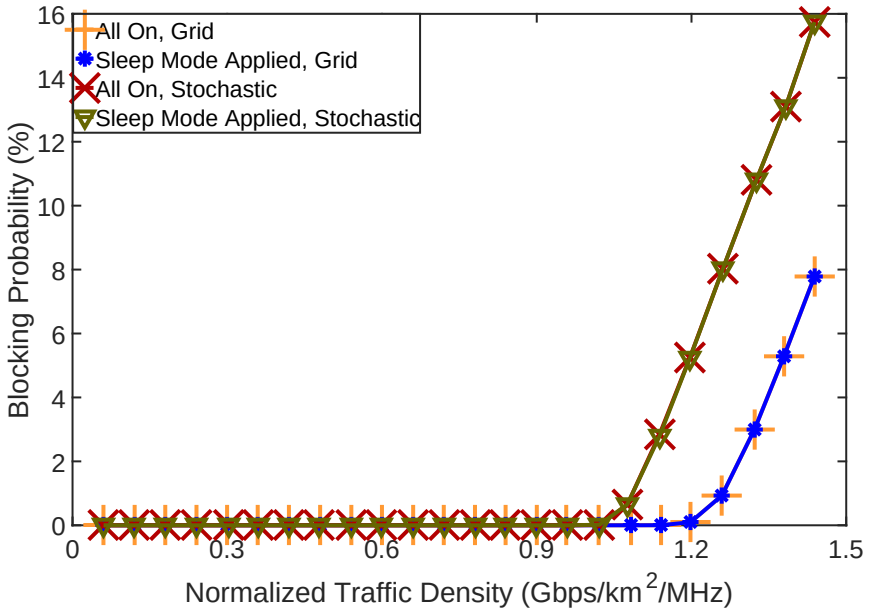
5.5.1 Performance Comparison

To understand the difference between the hexagonal grid model and the stochastic MHPP model in BS spatial modelling, the system performance with and without the ATP based sleep mode operation on the ideal and non-ideal eNB layouts is compared. The hexagonal grid model used is of the same configurations, with 10.7 metres ISD and 91 grids forming 6 rings of hexagons. The MHPP is chosen to be realised on a square with a 95 metres side length. For the MHPP, λ_{p} is set to 20000/km^2 and δ is configured to 5 metres to also make eNB densities in two scenarios equivalent. The values are chosen to make the networks in two scenarios have the same area and the identical density of BSs. The rest of the system parameters and the simulation methods are the same as stated in Section 5.4.

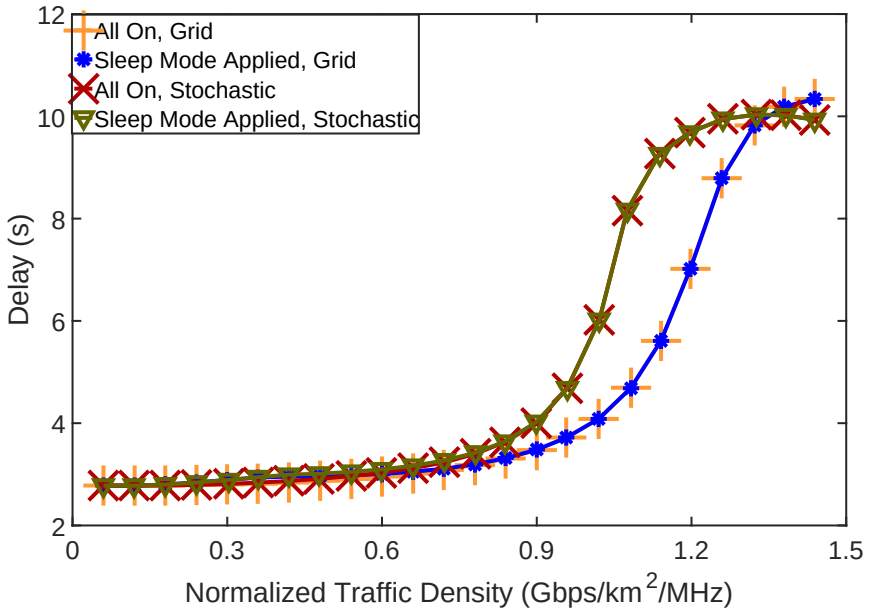
Delay and blocking probability presented in Fig. 5.5 show consistent changes, where the system performance on the MHPP model is worse than that on the hexagonal grid model. Interference is more severe due to the irregular placement of eNBs, yielding higher blocking probability and delay. On the other hand, the proposed ATP based sleep mode operation is shown to have no variations in tracing the system QoS without sleep mode operation. With the eNB densities adopted for both models, the same area power consumption can be observed in Fig. 5.6 for a system with all eNBs active. The ATP based sleep mode operation behaves similarly in reducing power consumption with a similar number of active eNBs for different traffic levels. In other words, the ATP based sleep mode operation algorithm is robust to various layouts of SCNs and shows a consistent performance in maintaining QoS and reducing power consumption.

5.5.2 Probabilistic Analysis

Consider that a SCN modelled with a MHPP Φ_{m} with intensity λ_{p} and a hard-core distance δ on $\mathcal{A}$ is implemented with the ATP based sleep mode operation. Intensity of Φ_{m}, λ_{m}, is then $\lambda_{\mathrm{p}}p$, where p is given in Equation (4.8). The sleep mode operation can be regarded as a second thinning process based on $\Phi_{\mathrm{m}} \cap \mathcal{A}$, during which the active eNBs are retained. The sleep mode operation strategy adopted at time t can be denoted as the retaining probability



(a)



(b)

Figure 5.5: QoS on the ideal and non-ideal eNB layouts against normalised traffic density. (a) Blocking probability. (b) Delay.

of the second thinning process as $\eta_t(A)$, which is a variable of location $A \in \mathcal{A}$. It essentially relates to the intensity of active eNBs, $\lambda_{\mathrm{m}}\eta_t(A)$. Therefore, without considering the energy consumption during transient states for simplicity, area power consumption given in Equation (3.14) can be written as

$$\overline{P_t^{\mathrm{N}}} = \int_{\mathcal{A}} \frac{\eta_t(A)\lambda_{\mathrm{m}}P_{\mathrm{B}} + \big(1-\eta_t(A)\big)\lambda_{\mathrm{m}}P_{\mathrm{S}}}{|\mathcal{A}|}\mathrm{d}A, \tag{5.6}$$

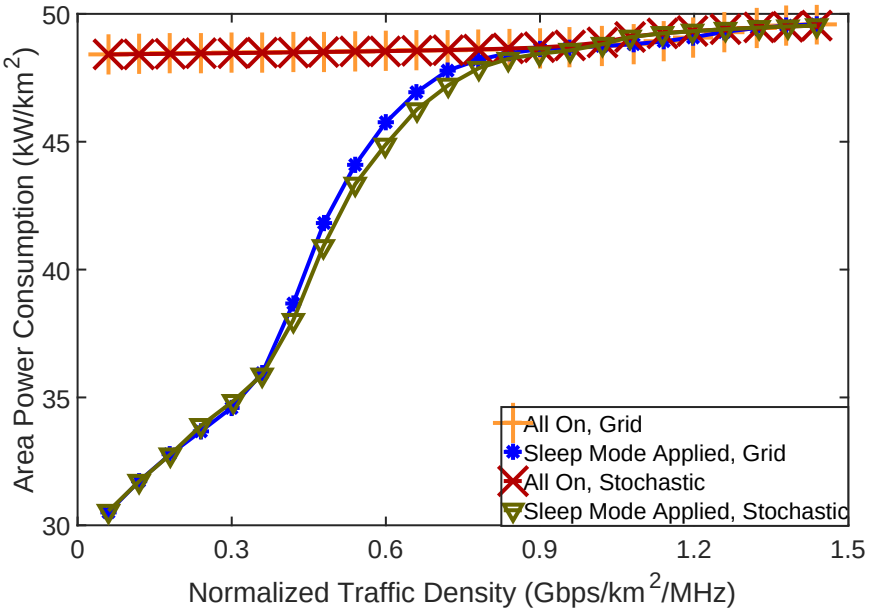


Figure 5.6: Area power consumption on the ideal and non-ideal eNB layouts against normalised traffic density.

where P_{B} and P_{S} are the eNB power consumption at the active mode and the sleep mode, respectively, as modelled in Chapter 3.

The ATP based sleep mode operation reduces $\overline{P_t^{\mathrm{N}}}$ by keeping $\eta_t(A)$ as low as possible without affecting QoS for every A since $P_{\mathrm{S}} < P_{\mathrm{B}}$. For a uniform traffic distribution investigated in this chapter, $\eta_t(A)$ can be approximated to be uniform versus location A and is thus denoted as $\overline{\eta_t}$. The ATP mechanism builds monitoring areas depending on the distances to the nearest eNBs considering the ideal distance case, which has a minor deviation from the practical implementation in terms of the system performance. Thus, $g_n(r)$, the derivative of $G_n(r)$ approximated in Section 4.4.2 and Section 4.4.3 should be utilised to understand how the system performance is affected by the ATP based sleep mode operation. As analysed in the performance evaluation, the trade-off between QoS and power consumption is the most critical. Hence, the relationship among one of the algorithm parameters, average delay $\overline{\varepsilon}$ and $\overline{\eta_t}$ is studied.

For a specific $B_i \in \mathcal{B}$, the probability that its nth nearest eNB B_i^n is its nearest active eNB at time t can be calculated as

$$\mathbf{P}\left(\arg \min_{B_j \in \mathcal{B}^{\mathrm{on}},\ j \neq i} D_{i,j} = B_i^n\right) = \overline{\eta_t}(1-\overline{\eta_t})^{n-1}, \tag{5.7}$$

The calculation from the perspective of sleep mode operation is equivalent to considering the second thinning process of Φ_{m} with identical retaining probabilities. As $g_n(r)$ captures the

probability that the distance from B_i to its nth nearest eNB in Φ_{m} is r, the probability that the distance from B_i to its nearest active eNB at time t is r can be obtained as

$$\mathbf{P}\Big(\min_{B_j\in\mathcal{B}^{\mathrm{on}},\ j\neq i} D_{i,j}=r,\ \arg\min_{B_j\in\mathcal{B}^{\mathrm{on}},\ j\neq i} D_{i,j}=B_i^n\Big)=g_n(r)\overline{\eta_t}(1-\overline{\eta_t})^{n-1}. \tag{5.8}$$

Denoting the mean interarrival time between two files as $\overline{\lambda_{\mathrm{f}}}$ using the traffic model introduced in Section 3.2.6, the expectation of the number of active links for a UE can thus be estimated by $\overline{\varepsilon}/\overline{\lambda_{\mathrm{f}}}$. Therefore, the expectation of the number of active links in a circle of radius r is equal to $\overline{\varepsilon}\lambda_{\mathrm{u}}\pi r^2/\overline{\lambda_{\mathrm{f}}}$, where λ_{u} is the user intensity. If the same ATP based sleep mode operation strategy is selected with a fixed W_{th}, $\overline{\eta_t}$ can be simplified to $\overline{\eta}$. Finally, the expected number of active links in a monitoring area is acquired by considering all the probabilities as

$$\frac{\sum_{n=1}^{\infty}\int_{\delta}^{\infty} g_n(r)\overline{\eta}(1-\overline{\eta})^{n-1}\overline{\varepsilon}\lambda_{\mathrm{u}}\pi r^2\mathrm{d}r}{\overline{\lambda_{\mathrm{f}}}}=\overline{|\mathcal{W}_i|}. \tag{5.9}$$

where $\overline{|\mathcal{W}_i|}$ is the average number of active links in the monitoring areas. The relationship is based on the assumption that the requirements for switching to sleep modes do not consider L_i, which usually meets the condition when $|\mathcal{W}_i|$ does at medium or high network traffic levels. With the assumption, $|\mathcal{W}_i|$ dropping below W_{th} may yield a new sleeping eNB and result in the potential increase in mean of $|\mathcal{W}_i|$, which is the opposite effect of an activation process. Hence, W_{th} can be adopted to estimate the expected number of active links in a monitoring area for low and medium traffic levels, where W_{th} maintains $|\mathcal{W}_i|$ by activations and deactivations.

From the probabilistic analysis, Equation (5.9) reveals the insight in ATP based sleep mode operation, where the relationship among the algorithm parameter $\overline{|\mathcal{W}_i|}$, $\overline{\varepsilon}$, λ_{f} and $\overline{\eta}$ is quantified. The metrics $\overline{\varepsilon}$ and $\overline{\eta}$ relate to delay and area power consumption, respectively. λ_{f} relates to traffic density, being the x-axis of the figures. The quantitative correlation can be used to predict any metric if the values of the others or their correlations are known. In Fig. 5.7, the results of the ATP based sleep mode operation ($W_{\mathrm{th}}=2$, $H=10\ s$) are used to predict $\overline{|\mathcal{W}_i|}$ for different λ_{f}, which are converted to traffic densities. It is worth mentioning that $\overline{|\mathcal{W}_i|}$ is around 2.5 at 0.8 Gbps/km^2/MHz, which is the traffic density where almost all eNBs are active as shown in Fig. 5.6. With $W_{\mathrm{th}}=2$, an activation occurs when $|\mathcal{W}_i|=3$ while a deactivation happens when $|\mathcal{W}_i|=2$. Their mean indicates $\overline{|\mathcal{W}_i|}$ for the traffic levels, where there are both activations and deactivations existing, i.e. not all eNBs are active. This matches the fact that $\overline{|\mathcal{W}_i|}$ keeps around 2 at low and high traffic levels, and crosses 2.5 when

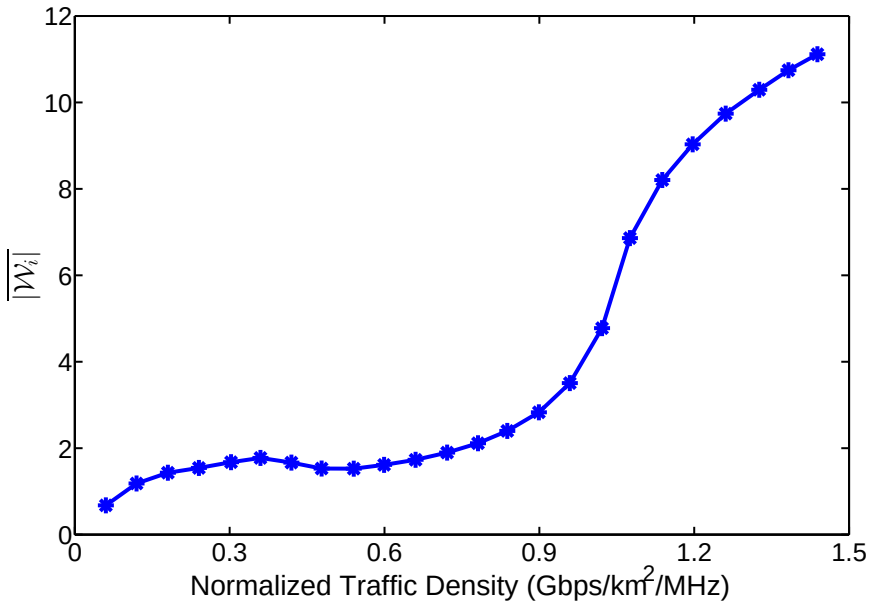


Figure 5.7: $\overline{|\mathcal{W}_i|}$ predicted by Equation (5.9).

almost all eNBs are active.

5.6 Conclusion

This chapter proposes a fast reaction, protocol-friendly, sleep mode operation based on adaptive traffic perception for distributed SCNs. The implementation of the algorithm would be practical with the current LTE system. It relies on adaptively changed monitoring areas of power forms to measure local traffic levels. Loads, the numbers of active links and the monitoring radii are exploited as the traffic information, based on which all the eNBs made the decisions of switching independently. It is demonstrated to be able to maintain QoS while reducing area power consumption, around 37% at a low traffic level. It is also analysed on a stochastic model for spatial modelling of eNBs with random placement. Its performance is compared with the one on the hexagonal grid based model, showing a consist performance in QoS maintenance and power reduction, which reveals its robustness to different SCN topologies.

Chapter 6

Traffic-Aware Cell Management

Contents

6.1 Introduction

As reviewed in Section 2.4.3, sleep mode operation algorithms can be categorised regarding the types of the network elements which control the operation. The ATP based sleep mode

operation algorithm proposed in Chapter 5 distributes the decision making and controlling functions to all the network node, including the sleeping ones. As a robust algorithm, it achieves the objective of reducing the overall power consumption while maintaining the QoS, however, has a limitation in the requirements of the functionalities of sleep modes. It requires sleeping eNBs to perform some computation, including comparing the receiving power of indicator or reference signals with the monitoring radii in power forms and counting the active links in their monitoring areas. Although the computation responsibilities are light considering the computation complexity, it needs more active hardware modules in sleep modes, especially the computation units, which generate more power consumption. As a result, the reduction in the overall power consumption of SCNs may be attenuated by higher power consumption of sleep modes.

The computation unit of an eNB is better to be deactivated in a sleep mode if the computation tasks consume non-negligible power. Then, a sleeping eNB does nothing unless it receives an activation message from other eNBs, which only needs the function of receiving and no extra computations. In this case, different formats of traffic information are required to provide active eNBs enough knowledge to make decisions for the sleeping ones. There has already been some research carried out to suit this preferred distributed manner. For example in [56], an algorithm is presented to control sleep modes in conventional cellular networks, which needs load information and coverage information exchange among neighbour BSs. A more consolidated scheme is proposed and analysed in [83], where a BS considers both the load information of its neighbour BSs and the handover information of the associated UEs to make switch-off decisions. Activation decisions are made by neighbour BSs based on their load conditions. These schemes require no computations at sleeping BSs, which allows a sleep mode to consume less power. However, all these schemes mentioned need some form of information exchange among neighbour BSs, such as load information, which yields additional system overheads.

In addition to the extra system overheads, the other cost is on the UE side, which is the increase in the average uplink transmission power. With the application of sleep modes, there are fewer active eNBs in a network so that the average distance between UEs and their corresponding associated eNBs is longer. This drives UEs to increase the transmission power to compensate for the path loss, which reduces their battery lives. Although the SCNs with much smaller cell sizes have reduced the transmission power of UEs dramatically, and the increase in the UE transmission power is negligible compared with the significant

power reduction at the eNB side, it is essential to minimise this effect, which is more obvious considering the existence of hotspot areas. When designing a sleep mode control algorithm, active eNBs should be placed closer to UEs, yielding irregular cells in terms of shape and size to cope with the unbalanced traffic distribution. However, this is rarely considered in the literature.

This chapter proposes a Traffic-aware Cell Management (TACM) scheme to exploit different traffic information to realise no computation at sleeping eNBs, fewer system overheads and less UE transmission power by managing the actions of small cells. It achieves the objective by making active eNBs determine the best time for switching on-off. Direction of Arrival (DOA) information of the UEs is used as one of the decision making criteria. Its utilisation aims to reduce the associated system overheads, including the information exchange overhead, and to place active eNBs where the increase in the UE uplink transmission power can be minimised.

The remainder of the chapter is organised as follows. In Section 6.2, DOA related background information is provided. The details of the proposed TACM algorithm including DOA classification, weight update, cell management are described in Section 6.3 and Section 6.4. In Section 6.5, the TACM algorithm is evaluated, analysed and compared with a baseline scheme. Finally, conclusions are drawn in Section 6.6.

6.2 DOA Related

The motivation for adopting DOA information is based on the fact that DOA estimation has been studied over the past few decades and is widely used for smart antenna beamforming, which is expected to deliver increased performance benefits in future networks. With the spread of multiple antenna systems, the availability of DOA information can be guaranteed and used for sleep mode control in the SCNs investigated. Conventionally, DOA can be estimated by uniform sensor arrays as well as either isotropic or directional antenna arrays using various algorithms, e.g. MUSIC [84], which is compared and analysed with other algorithms in [85]. The experimental work done is as rich as the theoretical algorithms. For instance in [86], MUSIC is evaluated with an antenna array of 6 parasitic elements, which achieves an error of less than one degree under the conditions of a 20 dB signal-to-noise ratio (SNR), 1000 snapshots and a 500 kHz sampling rate. In spite of this, attention still has to be paid to the resolution property of DOA estimation which depends upon the SNR, the

number of elements in the array, the number of snapshots, the array geometry and so on [85]. Although, DOA estimation is more effective for LOS paths, research is also ongoing to deal with the influence of NLOS paths [87, 88]. In the SCN scenario investigated in this thesis, the transmission path between a UE and a BS is usually short and is more likely to be LOS so that DOA estimation can potentially have high accuracy. The DOA estimation error is also quantified to investigate its effects on the TACM performance as introduced in later contents.

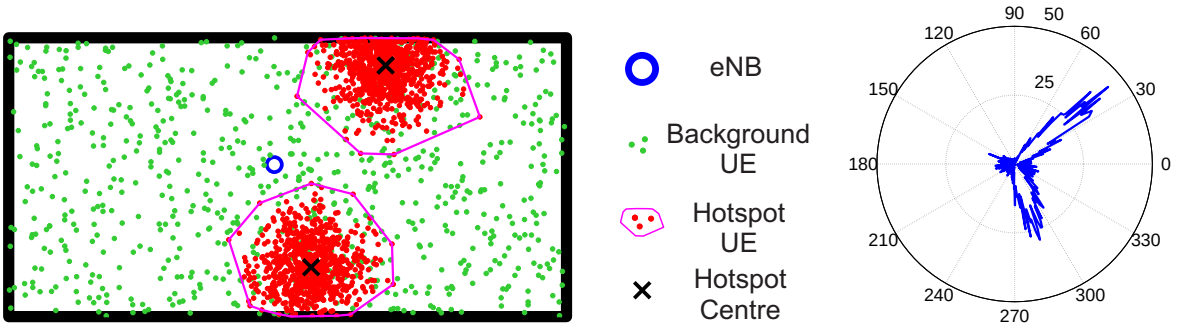
6.3 Traffic Awareness

As aforementioned, as one of the objectives, it is necessary to consider the placement of active eNBs when designing a sleep mode operation scheme. Due to the existence of hotspot areas, the key point is to ensure that a SCN always provides just enough radio resources in appropriate areas in pace with the temporal and spatial traffic variations. Thus, it is vital to locate more active eNBs delivering services at places with high service demands and irregular cells can be created to cover areas of different traffic levels in an accommodative manner. In this way, UEs are prevented from connecting far away causing severe interference and wasting transmission power.

In the TACM algorithm, the concept of traffic awareness is introduced to provide an eNB knowledge of the relative directions to hotspot areas based on the traffic information collected. It relies on DOA estimation in an observation process, after which DOAs are classified and distributed weight memories are created at all eNBs. The memory of an eNB is accumulatively updated through its past experience to acquire the awareness of the traffic distribution.

6.3.1 Observation

This observation process requires only active eNBs to estimate the DOAs of their admitted UEs. When UEs are preparing for file transmissions, DOA estimations are performed based on control signals containing different UE identifications. As noted earlier, DOA estimation is assumed to be ideal. The example in Fig. 6.1 (a) assumes that the eNB estimates the DOAs of all the UEs, each of which has one file arrival. The counted number of DOAs with 1 degree resolution as an example is plotted in Fig. 6.1 (b). It is obvious that the estimation of DOAs yields an angular distribution of the traffic perceived by the eNB and indicates the



(a) (b)

Figure 6.1: DOA estimation. (a) The eNB estimates the DOAs of all the UEs. (b) The angular distribution of the perceived traffic. The traffic level is the number of DOAs in the corresponding directions with 1 degree resolution.

directions of higher traffic levels.

6.3.2 Memory Update

The memory update process is executed every time a new DOA estimation is performed. The DOA is further classified and the corresponding metrics are updated. In this process, an eNB can determine the directions of the areas with high traffic levels and take them as part of the decision making criteria.

The eNBs in a SCN network are denoted by a set $\mathcal{B} = \{B_i\}$ $(i \in \mathbb{N}^*)$ and the neighbour eNBs of a eNB B_i are represented by a set $\mathcal{N}_i = \{N_{i,j}\}$, where $j \in \mathbb{N}^*$. To quantify the captured DOAs for each B_i, a weight memory ($\mathcal{W}_i = \{W_{i,j}\}$) is needed and each memory unit in the weight memory is mapped to one of the neighbour eNBs $N_{i,j} \in \mathcal{N}_i$ of B_i as

$$\forall W_{i,j} \in \mathcal{W}_i : W_{i,j} \mapsto N_{i,j}. \tag{6.1}$$

When B_i creates its own NCL, the angle of each $N_{i,j}$ relative to B_i, $D_{i,j} \in [0,360)$ is acquired and recorded in degrees. Then for DOA classification, an interval $\mathcal{I}_{i,j}$ for each $W_{i,j}$ representing a sector with its angular bisector threading $N_{i,j}$ is generated and its centrally symmetric sector represented by an interval $\mathcal{I}^*_{i,j}$ is also created as

$$\mathcal{I}_{i,j} = (D_{i,j} - \frac{\Delta\theta}{2}, D_{i,j} + \frac{\Delta\theta}{2}), \tag{6.2}$$

$$\mathcal{I}^*_{i,j} = (D_{i,j} - \frac{\Delta\theta}{2} + 180, D_{i,j} + \frac{\Delta\theta}{2} + 180), \tag{6.3}$$

where $\Delta\theta \in (0,180]$ is a predefined parameter in degree representing the interval length. In

this way, B_i updates a weight $W_{i,j}$ for every neighbour eNB $N_{i,j}$, reflecting the traffic level in that direction relative to B_i. Like the example given in Fig. 6.2, B_3 denotes its first neighbour eNB B_2 as $N_{3,1}$ and creates an interval $\mathcal{I}_{3,1}$ for it. The angle of B_2 relative to B_3 is 354°, then $\mathcal{I}_{3,1}$ becomes (324°,384°) according to Equation (6.2) with $\mathcal{I}^*_{3,1}$ being (504°,564°) according to Equation (6.3) if $\frac{\Delta\theta}{2}$ is 30°.

When a UE is admitted by B_i for data transmission, the memory size ψ_i, defined as the total entries of UEs, is increased by 1 regardless of their DOAs. Meanwhile, the estimated DOA of the UE is represented by a relative angle $\alpha \in [0,360)$. Afterwards, the entry of this UE is classified into the intervals which it belongs to and the corresponding weights are updated respectively as

$$\psi_i^{(t+1)} = \psi_i^{(t)} + 1, \tag{6.4}$$

$$\begin{aligned} &\exists k \in \{0,\pm 1\} : (\alpha + 360k) \in \mathcal{I}_{i,j} \\ &\qquad \Rightarrow W_{i,j}^{(t+1)} = W_{i,j}^{(t)} + 1, \end{aligned} \tag{6.5a}$$

$$\begin{aligned} &\exists k \in \{0,\pm 1\} : (\alpha + 360k) \in \mathcal{I}^*_{i,j} \\ &\qquad \Rightarrow W_{i,j}^{(t+1)} = W_{i,j}^{(t)} - 1, \end{aligned} \tag{6.5b}$$

$$\begin{aligned} &\forall k \in \{0,\pm 1\} : (\alpha + 360k) \notin (\mathcal{I}_{i,j} \cup \mathcal{I}^*_{i,j}) \\ &\qquad \Rightarrow W_{i,j}^{(t+1)} = W_{i,j}^{(t)}, \end{aligned} \tag{6.5c}$$

where the superscript (t) is the prior time step of $(t+1)$. The $360k$ terms in Equation (6.5) exist in case that the ranges of $\mathcal{I}_{i,j}$ or $\mathcal{I}^*_{i,j}$ exceed 360° or fall below 0° ($\alpha + 360k$ is equal to α). As a result, the corresponding weight is increased by 1 if a DOA is classified in $\mathcal{I}_{i,j}$ or decreased by 1 if a DOA is classified in $\mathcal{I}^*_{i,j}$. Multiple weights may be updated for one entry if it is classified into multiple overlapping intervals.

In one case, the weight $W_{i,j}$ is relatively high if more DOAs are classified into the interval $\mathcal{I}_{i,j}$ indicating that the traffic level in the corresponding direction range $\mathcal{I}_{i,j}$ is high. In the other case, fewer DOAs classified into the interval $\mathcal{I}^*_{i,j}$ indicate less traffic offered in the opposite direction range of the one pointing to $N_{i,j}$, which also yields a higher weight $W_{i,j}$. Implicitly, as active eNBs should stay close to where there are high traffic levels, or stay far away from areas with low traffic levels, a relatively higher weight $W_{i,j}$ always implies that the corresponding direction range $\mathcal{I}_{i,j}$ points to an area where active eNBs should be located.

An example of DOA estimation and weight update is given in Fig. 6.2. The newly arriving UE_1, UE_3 and UE_4 are classified to $\mathcal{I}_{1,1}$, $\mathcal{I}_{3,1}$ and $\mathcal{I}^*_{3,1}$ according to the relative angles of the

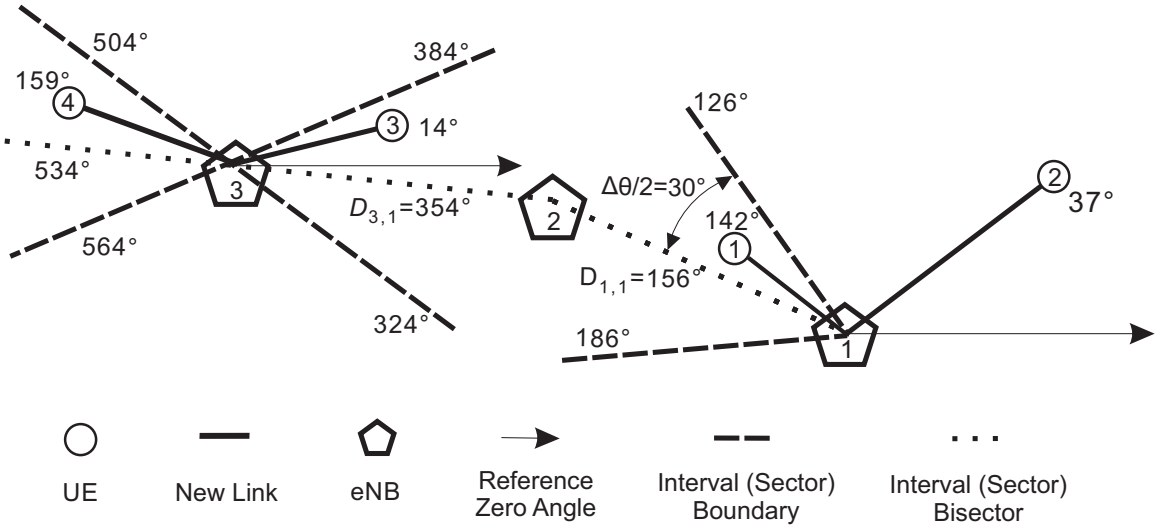


Figure 6.2: An example of interval creation, DOA classification and weight update. B_2 is denoted as $N_{1,1}$ and $N_{3,1}$, indicating that it is the first neighbour eNB of B_1 and B_3, respectively. The intervals for neighbour eNBs are built. For instance, the angle of B_2 relative to B_3 is 354°, then $\mathcal{I}_{3,1}$ is (324°,384°) according to Equation (6.2) with $\mathcal{I}^*_{3,1}$ being (504°,564°) according to Equation (6.3) assuming $\frac{\Delta\theta}{2}$ to be 30°. UE_1 and UE_2 are admitted by B_1 while UE_3 and UE_4 are admitted by B_2. Equation (6.5) is used for classification. UE_1 is classified to $\mathcal{I}_{1,1}$ of B_1 while UE_3 and UE_4 are classified to $\mathcal{I}_{3,1}$ and $\mathcal{I}^*_{3,1}$ according to the relative angles of the DOAs and the reference zero angles. UE_3 with 14° DOA is classified to $\mathcal{I}_{3,1}$, (324°,384°), as 14 plus 360 is 374 with $k = 1$ from Equation (6.5b).

DOAs and the reference zero angles. This is processed based on Equation (6.5). After that at time step $(t+1)$, the weight memory unit $W^{(t)}_{1,1}$ is increased by 1. $W^{(t)}_{3,1}$ remains the same because of one step of increase contributed by UE_3 and one step of decrease contributed by UE_4. However, UE_2 is not classified into any intervals and therefore no corresponding weights are changed for this entry. On the other hand, $\psi^{(t)}_1$ and $\psi^{(t)}_3$ are both increased by 2 at time step $(t+1)$ according to Equation (6.4).

For a certain eNB, its weights should be normalised to the overall traffic supported by it if it is required to know the relative traffic levels (irrespective of the absolute traffic level) in different direction ranges pointing to its neighbour eNBs. For this purpose, weight bias $\beta_{i,j}$ is defined as

$$\beta^{(t)}_{i,j} = \frac{W^{(t)}_{i,j}}{\psi^{(t)}_i}. \tag{6.6}$$

It is easy to find out that $\beta_{i,j}$ is also mapped to $N_{i,j}$ because of the association with $W_{i,j}$. $\beta_{i,j}$ also stands as a direction indicator of hotspot areas like $W_{i,j}$, but with normalisation reflecting the relative traffic level to a direction.

From the previous explanation, it is known that the direction range $\mathcal{I}_{i,j}$ with a corre-

sponding higher weight $W_{i,j}$ points to the area where more traffic arrives. By normalising $W_{i,j}$ to the memory size ψ_i, the relative trends of service demands in areas in different directions are quantified by $\beta_{i,j}$. A higher value of bias signifies the direction where the cell supporter should be. To this point, the traffic awareness is formed, based on which proper actions can be taken to place active eNBs close to hotspot areas.

As the state transitions of eNBs change the coverage areas of the active eNBs, this may give rise to inconsistency between the historical memory and the existing traffic loaded by an eNB. To deal with this, the memory of an eNB should be updated after the action has been taken, making the memory more reliable for the next action to be taken. This is described in the following section.

6.4 Cell Management

With the traffic awareness, an active eNB has the knowledge of the relative directions to hotspot areas thanks to the exploitation of DOA information as a part of traffic information. It thus has the ability to manage the neighbour sleeping eNBs, which can be implemented with sleep modes with lower power consumption. An activation decision for a sleeping eNB can be made by another active eNB only based on its own historical DOA perception and its own average load estimation, which is the other part of traffic information utilised. The active eNB does not require the load information of neighbour eNBs and therefore mitigates the information exchange overhead. With the joint efforts of all eNBs in a SCN, the SCN can manage its cell configurations according to the varying environment, achieving the objectives of no computation at sleeping eNBs, fewer system overheads and less UE transmission power as mentioned in the introduction. The cell management is introduced in detail in this section.

6.4.1 Cell Division

Cell division helps an overloaded eNB to activate one of its sleeping neighbour eNBs in the desired direction. It is triggered based on the local traffic levels, after which the updated memories help eNBs determine which neighbour eNB to be activated. Whenever the load L_i of B_i reaches a division threshold $DivL_{\text{th}}$ for a continuous hysteresis duration H_{div} timed by the hysteresis timer, cell division is triggered. The hysteresis duration H_{div} is to prevent B_i from being perturbed by temporary short-term traffic variations which may easily make the load of B_i exceed the division threshold, $DivL_{\text{th}}$. This reduces the ping-pong effect

and therefore reduces the overall switching on-off frequency. As mentioned before, through cumulative DOA estimation and weight update from the past experience of B_i, a higher value of $W_{i,j}$ indicates that there is a high traffic level in the direction to $N_{i,j}$. After cell division is triggered, B_i selects and activates only one $N_{i,j} \in \mathcal{N}_i^{\text{off}}$ mapped to the largest weight, where $\mathcal{N}_i^{\text{off}}$ is the neighbour sleeping eNB set of B_i. As the NCL is already initialised after deployment, B_i sends an activation message via the air interface to its selected neighbouring eNB. A sleeping neighbour eNB has a minimum message receiving capability for control purposes and is activated when receiving such messages. It is worth mentioning that such messages are only transmitted when an activation occurs and do not require the eNBs to frequently communicate with each other. Then, the activated eNB recovers normal operation and starts to broadcast reference signals to serve UEs again.

To provide sufficient resources by having more cells in a SCN, cell division occurs by activating one of the neighbour sleeping eNBs of an overloaded eNB. Then, B_i resets every element in $\mathcal{W}_i$ and ψ_i to zero, and empties the hysteresis timer. As shown in the example in Fig. 6.3 (a), B_1 is overloaded by its local traffic load and perceives a high traffic level from its right. It activates its right neighbour eNB B_3 to support the high traffic level. Consequently, B_1 and B_3 adjust their coverage areas following a certain UE association policy and more resources are delivered to the previous overloaded local area.

6.4.2 Cell Migration

Due to the lack of logical interfaces between non-neighbour eNBs in a distributed SCN, only neighbour eNBs can be activated in cell division to satisfy the neighbour areas in demand. Therefore, eNBs can be activated in the required directions but not at the expected distances. Moreover, SCNs only have spatial traffic variations sometimes so that only the locations of active eNBs have to be adjusted and the overall number of active eNBs may remain the same. By guaranteeing the correct locations of active eNBs, a combination of both desirable directions and distances, the average distance between UEs and eNBs and the resulting average UE uplink transmission power can be minimised.

Motivated by the above requirements, cell migration is designed. In cell migration, if an eNB perceives a relatively high traffic level in the direction interval mapped to a neighbour eNB, it will activate this neighbour eNB and switch off itself. In this way, the newly activated eNB is located nearer to the area with a higher traffic level. The migration process can be triggered taking $\beta_{i,j}$ into consideration when the local traffic level in a cell is not high, but of

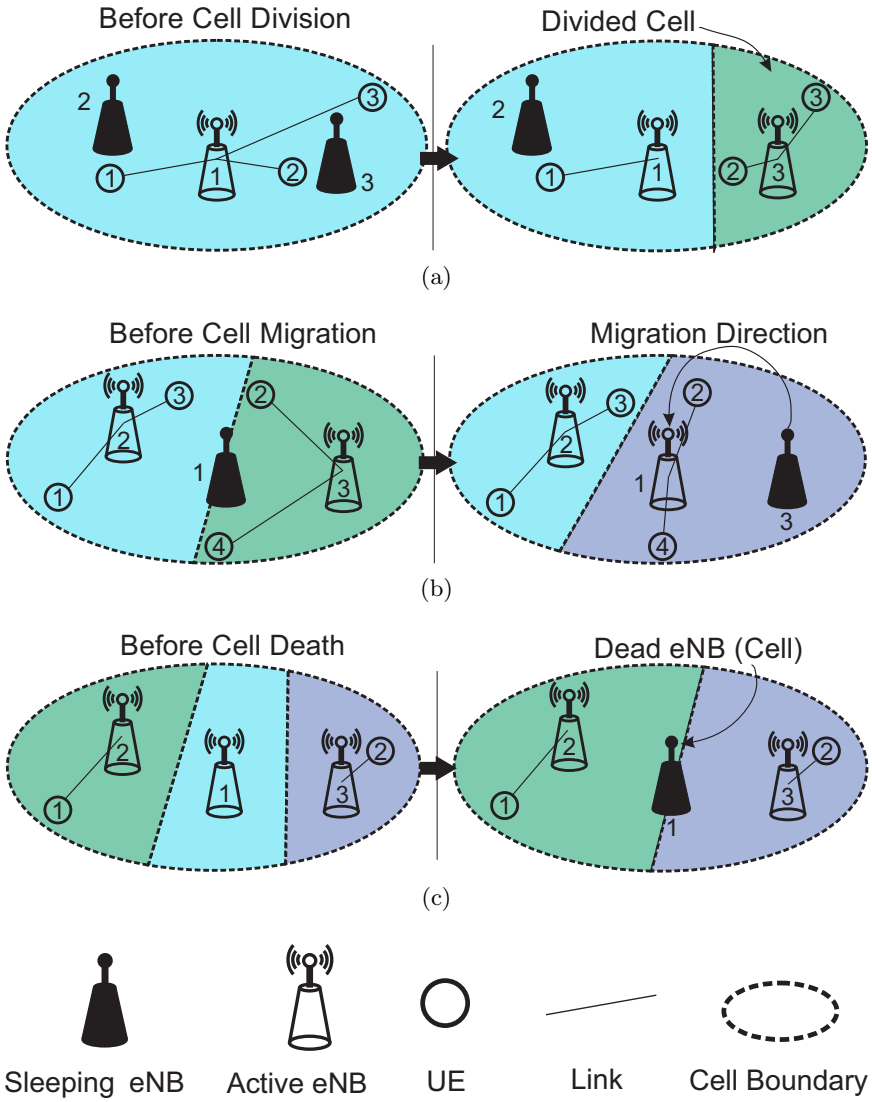


Figure 6.3: Examples of the action moments in the TACM algorithm. (a) Cell division: B_1 activates B_3, dividing the bigger cell into two small cells when B_1 is overloaded and perceives that the hotspot area is to the right. (b) Cell migration: B_3 perceives that the angular traffic distribution is biased to the left, then it activates B_1 and deactivates itself, moving the cell supporter closer to the hotspot area. (c) Cell death: B_1 switches to sleep mode when it has little load, killing the cell it served.

a biased angular distribution. More specifically, migration of a cell supported by B_i happens when there is a $\beta_{i,j}$ exceeding a bias threshold β_{th} and $N_{i,j} \in \mathcal{N}_i^{\text{off}}$. Meanwhile, the memory size ψ_i should be over a threshold ψ_{th} ensuring that a sufficiently large number of DOAs and a reliable past experience for reasoning are captured. After satisfying the conditions, the corresponding mapped neighbour eNB $N_{i,j}$ is activated by B_i. If there are multiple $\beta_{i,j}$ exceeding the bias threshold β_{th} with their mapped eNBs currently in sleep modes, B_i chooses the neighbour eNB with the largest $\beta_{i,j}$. Then B_i resets every element in $\mathcal{W}_i$ and ψ_i, and the hysteresis timer to zero. B_i switches to sleep mode afterwards.

After a migration step, the newly activated eNB supporting the cell may perceive less biased traffic because of the changes in DOAs. In another case, the migration process may have several steps until a closest eNB relative to the hotspot area is activated. However, finally, due to less biased traffic perceived by the newly activated eNB, the cell migration chain will stop and the network becomes stable again. In the example given in Fig. 6.3 (b), a weight bias of B_3, which is mapped to its neighbour eNB B_1, is over the threshold, indicating the hotspot area is to its left. B_3 activates B_1 and deactivates itself, making the active eNBs located nearer to the area with a higher traffic level. The cell migration and the changed DOAs of UE_2 and UE_4 reduce the bias of the traffic distribution relative to the new cell supporter, B_1. This forms a feedback loop from the network adaptations and the gradual migration process is terminated until angularly uniform traffic is perceived by the new cell supporter.

6.4.3 Cell Death

To provide just enough resources in a SCN, fewer eNBs should be active when the overall traffic level is low. Cell death enables eNBs with low loads to be switched off in order to reduce the overall network power consumption. As an inverse process of cell division, when the load L_i of B_i is not greater than a death threshold $DieL_{\text{th}}$ for a continuous hysteresis duration H_{die} timed by the hysteresis timer, B_i switches itself to sleep mode, removing the cell supported by it. Similarly, H_{die} exists to prevent perturbation from temporary short-term traffic variations. With $DieL_{\text{th}}$, it can be chosen as a pair for cell death when considering $DivL_{\text{th}}$ and H_{div} as a pair of parameters for cell division. After cell death, B_i resets every element in $\mathcal{W}_i$ and ψ_i, and clears the hysteresis timer to zero before switching to sleep mode. Before switching to sleep mode, B_i broadcasts a sleeping message. Any eNBs receiving such message will mark it as in sleep mode if it is in their NCLs. Fig. 6.3 (c) shows an example of cell death, in which B_1 is of a low load, and deactivates itself, reducing the overall network power consumption. It is worth mentioning that the death of a cell may also lead to cell migration owning to the biased traffic perceived by other eNBs.

6.4.4 Algorithm Summary

As a result of the mechanism introduced, active eNBs observe the SCN situation by estimating DOAs and accumulatively update their respective memories, which quantify the historical

experiences of the eNBs. Network adaptation decisions are made in a distributed way and each eNB reasons the action to be taken independently without exchanging load information among neighbour eNBs, which mitigates the system overheads and becomes one of the virtues of the algorithm. The collection of traffic information is only undertaken by active eNBs, relaxing the responsibilities of the sleep ones, allowing lower power consumption at sleep modes. Moreover, the existence of hysteresis timers reduces the ping-pong effect caused by temporary short-term traffic variations, resulting in a lower switching on-off frequency, which can potentially reduce the energy consumption of transient states. Another advantage of TACM is contributed by the directional division and migration, which move active eNBs closer to hotspot areas with higher service demands and therefore reduce the UE average transmission power.

The TACM algorithm is summarised in the pseudo code as follows.

1: **for** a $B_i \in \mathcal{B}^{\text{on}}$, B_i **do**
2: Estimate DOA of the admitted UE with service request
3: **if** $\exists k \in \{0, \pm 1\} : (\alpha + 360k) \in \mathcal{I}_{i,j}$ **then**
4: $W_{i,j}^{(t+1)} \leftarrow W_{i,j}^{(t)} + 1$
5: **else if** $\exists k \in \{0, \pm 1\} : (\alpha + 360k) \in \mathcal{I}_{i,j}^{*}$ **then**
6: $W_{i,j}^{(t+1)} \leftarrow W_{i,j}^{(t)} - 1$
7: **else**
8: $W_{i,j}^{(t+1)} \leftarrow W_{i,j}^{(t)}$
9: **end if**
10: $\psi_i^{(t+1)} \leftarrow \psi_i^{(t)} + 1$
11: $\beta_{i,j}^{(t+1)} \leftarrow \frac{W_{i,j}^{(t+1)}}{\psi_i^{(t+1)}}$
12:
13: Cell Division (for $\mathcal{B}^{\text{on}}$):
14: **if** $L_i \geq DivL_{\text{th}}$ continuously for H_{div} **then**
15: **if** $\mathcal{N}_i^{\text{off}} \neq \varnothing$ **then**
16: Find $N_{i,aim} \leftarrow \arg\max_{N_{i,j} \in \mathcal{N}_i^{\text{off}}} W_{i,j}^{(t+1)}$
17: Activate $N_{i,aim}$
18: B_i resets every element in $\mathcal{W}_i$ and ψ_i, and the hysteresis timer to zero which initialises the memory
19: **else**
20: B_i resets the hysteresis timer to zero

end if
end if

Cell Migration (for $\mathcal{B}^{\text{on}}$):
if $(\psi_i^{(t+1)} \geq \psi_{\text{th}}) \wedge (\mathcal{N}_i^{\text{off}} \neq \varnothing)$ **then**
Find $\beta_{i,\max}^{(t+1)} \leftarrow \max_{N_{i,j} \in \mathcal{N}_i^{\text{off}}} \beta_{i,j}^{(t+1)}$
if $\beta_{i,\max}^{(t+1)} > \beta_{\text{th}}$ **then**
Find $N_{i,\text{aim}} \leftarrow \arg\max_{N_{i,j} \in \mathcal{N}_i^{\text{off}}} \beta_{i,j}^{(t+1)}$
B_i activates $N_{i,\text{aim}}$
B_i resets every element in $\mathcal{W}_i$ and ψ_i, and the hysteresis timer to zero which initialises the memory
B_i switches to sleep mode
end if
end if

Cell Death:
if $L_i \leq DieL_{\text{th}}$ continuously for H_{die} **then**
B_i resets every element in $\mathcal{W}_i$ and ψ_i, and the hysteresis timer to zero which initialises the memory
B_i switches to sleep mode
end if
end for

Cell Division and Migration (for $\mathcal{B}^{\text{off}}$):
for a $B_i \in \mathcal{B}^{\text{off}}$, B_i **do**
if B_i receives activation message from B_j where $B_i \in \mathcal{N}_j$ **then**
B_i switches to active mode
end if
end for

6.5 Performance Evaluation

With the models introduced in Chapter 3, the performance of the proposed TACM algorithm is evaluated in this section. The SCN is modelled on a square area of 100 metres side length

with non-ideal spatial distributions, where λ_p of the MHPP is set to 20000/km^2 and δ is configured to 5 metres. The eNB power consumption model is adopted for the evaluation of network power consumption.

It is also compared with a consolidated baseline scheme of sleep mode control proposed in [83] to demonstrate the advantages of the TACM algorithm over the state-of-art. The comparative baseline scheme is transplanted to suit the self-organised SCN with the random topology. To achieve the best energy saving, the load information of neighbour eNBs and the handover information coming from associated UEs are needed in the comparative scheme so that a notion of network impact can be defined and calculated. It requires active eNBs to obtain load information from their neighbour eNBs, which needs frequent load information exchange. In the scheme, an eNB determines to switch off when the network impact is below a level computed by a load threshold minus a half of a hysteresis margin, where the load threshold is set to 12% and the hysteresis margin is 16% to mitigate the inefficient switching on-off [83]. An eNB decides to activate a neighbour sleeping eNB when the load of itself exceeds half of the hysteresis margin plus its corresponding recorded load. The values are chosen to achieve the reduction in inefficient switching on-off to some extent. The proposed TACM algorithm and the baseline scheme are made to provide similar QoS as the system with no sleep mode operation for a fair comparison. More details about the baseline scheme can be found in [83].

6.5.1 System Overhead Comparison

Two aspects of system overheads generated by the TACM algorithm and the baseline algorithm are compared. As exchanging information among eNBs is always accompanied by the requirement for additional signalling, including information messages themselves and control signals, the frequency of the load information transmission is quantified. Additionally, the switching on-off frequency is also investigated since state transitions of an eNB take time and may consume extra energy. The state transitions of an eNB also compel other eNBs around to change their operating parameters, creating extra system overheads.

Load Information Transmission Overhead

As load information acquirement among neighbour eNBs is assumed in the baseline scheme, the frequency of load information transmission is quantified by assuming that an eNB sends

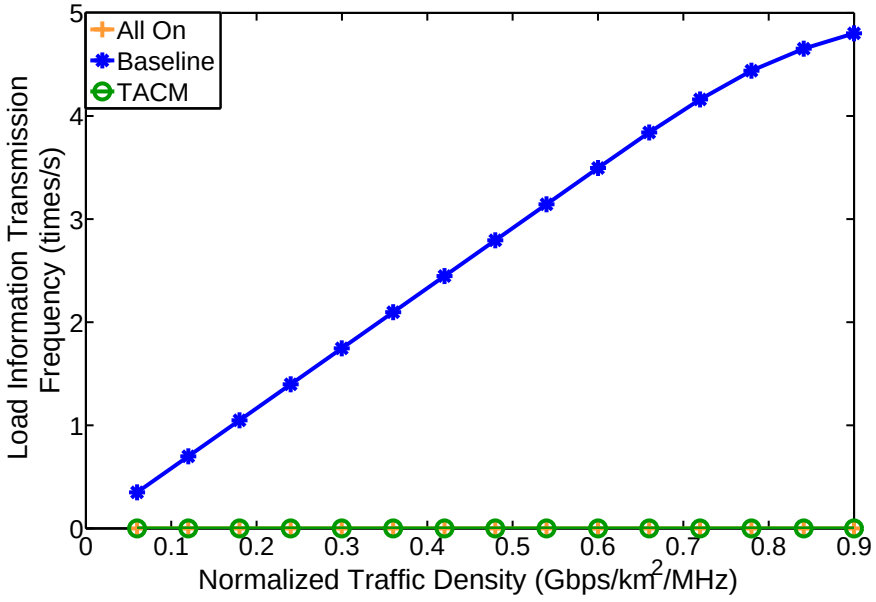
the load message whenever its load is varied and sending a message from an eNB to a neighbour eNB is counted as one process of load information transmission. In Fig. 6.4 (a), the average frequency of load information transmission is plotted against the average network traffic density, where the curve of the baseline algorithm shows an increasing trend because there are more load variations at an eNB when the overall traffic level increases. Obviously, the TACM algorithm does not require any load information exchange among neighbour eNBs so that the frequency is always zero, the same as the system without sleep mode operation. The TACM algorithm surpasses the baseline algorithm prominently in this aspect.

On-off Switching Overhead

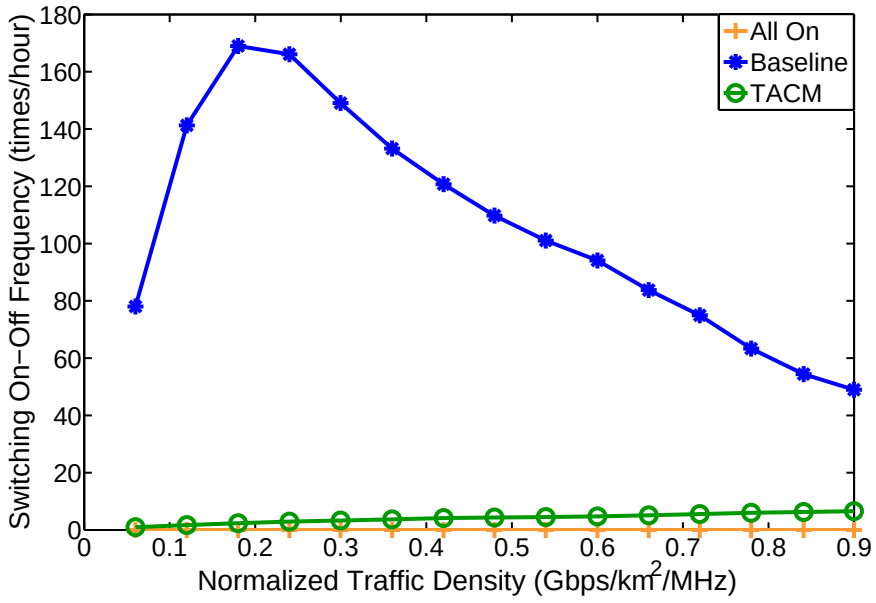
Fig. 6.4 (b) shows the average frequency of switching on-off where either switching on or switching off is counted. Although all sleep mode control schemes produce some system overheads due to the switching of eNBs, it can be reduced if an algorithm is comprehensively better. As can be seen from the simulation results, even with the application of the hysteresis margin in the baseline algorithm, the switching on-off frequency is very large, especially at medium traffic densities. At low traffic densities, fewer active eNBs can be switched off while fewer sleeping eNBs can be switched on at high traffic densities, both of which alleviate the switching on-off frequency to some extent. At medium traffic densities, the instant load can easily exceeds or falls below certain levels without averaging the load of an eNB over a period of time. With hysteresis durations predefined in the TACM algorithm, eNBs monitor their loads and make sure they are continuously above or below the corresponding thresholds in the hysteresis durations. In this way, TACM algorithm achieves up to 72 times less switching frequency compared with the baseline scheme, which thereby mitigates the system overhead dramatically.

6.5.2 Power Comparison

As one of the major metrics concerned in the conventional research of sleep mode control, area power consumption of both algorithms is compared. Moreover, as power control is involved in the system, the increase of UE transmission power on account of fewer active eNBs in the SCN along with sleep mode operation should be minimised, which is also investigated and illustrated.



(a)



(b)

Figure 6.4: Comparisons of the TACM algorithm and the baseline algorithm on system overheads. (a) The average frequency of load information transmission against network traffic density. (b) The average frequency switching on-off against network traffic density.

Network Power Consumption

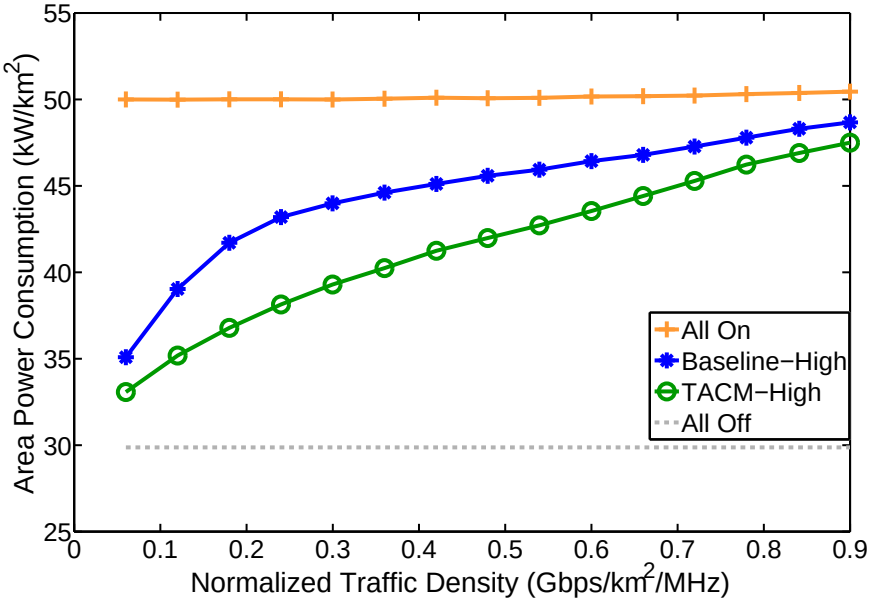
The network power consumption is investigated using area power consumption as in Fig. 6.5. As there is a trade-off between QoS and area power consumption, both schemes are first configured to ensure that the QoS is not degraded compared with no sleep mode application and then the power reduction is maximised. It is noticeable from Fig. 6.5 (a) that the area power

consumption reduced by the TACM algorithm is more than 34% (equivalent to 86% less overall active time on average) when the traffic level is below 0.1 Gbps/MHz/km^2 compared with no sleep mode control. Without the necessity of load information exchange, however, the TACM algorithm also reduces more area power consumption than the baseline algorithm, which is up to 81% more around 0.36 Gbps/MHz/km^2. This result is obtained by assuming a relatively high energy consumption during transient states (see power consumption model in Section 3.2.7). As can be seen from Fig. 6.5 (b), the average area power consumption of the baseline scheme is strongly affected by how much energy is consumed during transient states since its corresponding switching frequency is very high. Comparatively, the area power consumption generated by the TACM algorithm does not vary a lot versus the energy consumption of transient states (only the high level version is shown for clarity) since it mitigates the switching on-off overheads significantly while the high switching frequency of the baseline scheme may increase the overall area power consumption if the energy consumption of transient states are not negligible. Therefore, the proposed TACM algorithm also outperforms the baseline algorithm in terms of the switching frequency reduction and the resulting robustness to the potential cost of the energy consumption introduced by transient states.

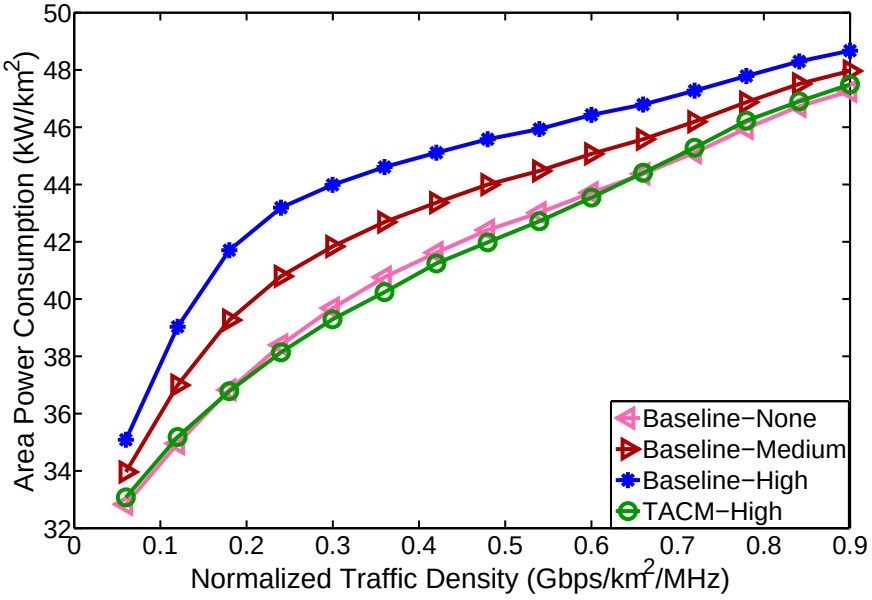
It is worth mentioning that the algorithm can still save some power even when the overall network load is heavy. This is because the spatial traffic distribution in the SCN is extremely heterogeneous in the investigated scenario with hotspot ratio γ equal to 0.8 so that some of the eNBs can be switched to sleep modes without affecting the overall QoS if the traffic levels are not high in their local areas. The area power consumption with all eNBs deactivated is denoted by the dashed line at the bottom of Fig. 6.5 (a), which actually represents the lower bound of area power consumption.

UE Transmission Power

Since the application of a sleep mode results in fewer active eNBs in the network, UEs increase their transmission power on average as a result of connecting to more distant eNBs during file transmission due to power control. Therefore, the average transmission power of UEs in this case is greater than that when sleep modes are not applied. To explore whether it is considerable compared to the power consumed by the eNBs, the average transmission power of UEs is plotted in Fig. 6.6. As observed in the figure, for each sleep mode operation algorithm, fewer active eNBs in the network correspond to higher average UE transmission power, implying a larger mean distance between eNBs and UEs. This signifies that there is



(a)



(b)

Figure 6.5: Comparisons of the TACM algorithm and the baseline algorithm on area power consumption. (a) Average area power consumption against network traffic density. (b) Average area power consumption assuming different energy consumptions during transient states against network traffic density.

trade-off between the reduction in area power consumption and the UE transmission power for a specific algorithm. However, although the transmission power of UEs is higher on average than the scheme with no sleep mode control when the traffic level is low, the energy radiated by all the UEs in the network during a certain period of time never exceeds 0.2% of the energy consumed by all the eNBs for both algorithms, meaning that the overall UE

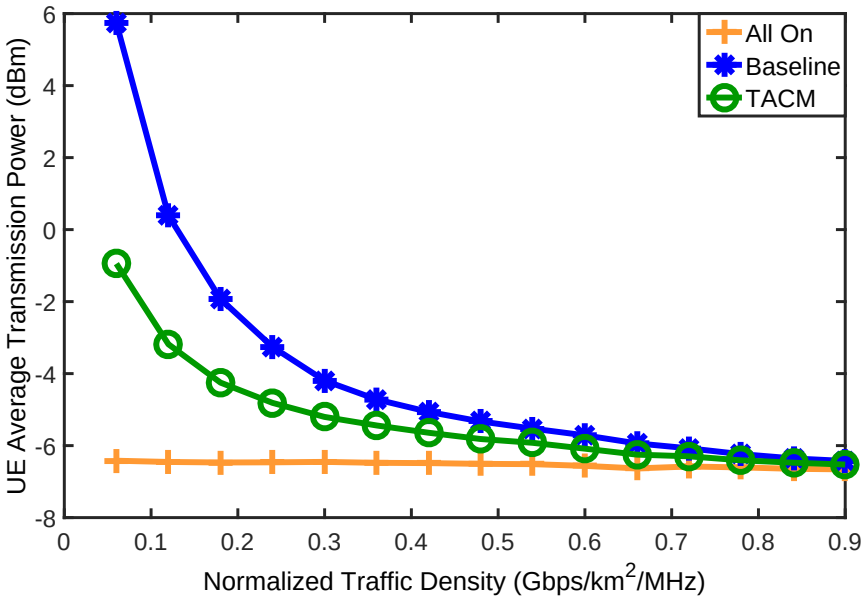


Figure 6.6: Comparisons of the TACM algorithm and the baseline algorithm on the average UE transmission power against network traffic density.

transmission power consumption is negligible relative to the amount of power consumption reduced by the sleep mode control algorithms.

On the other hand, considering the battery lifetime of the mobile devices, the increase of UE transmission power should be minimised by reducing the average distance between eNBs and UEs. Thanks to the directional activation and migration of the TACM algorithm, it is more likely to locate active eNBs at the places with high data service demands than the baseline algorithm. For this reason, the average transmission power of UEs can be 79% (computed with power in mW) less than that of the baseline algorithm at a low traffic level, which significantly improves the user experience.

6.5.3 Impact of Performance Related Parameters

The impact of the parameter variations of the TACM algorithm are demonstrated in this subsection. The tuning parameters of the TACM algorithm are varied to reveal the trade-off between QoS and area power reduction. The other critical parameters are also varied to observe how they influence the effectiveness of the TACM algorithm. The resulting impact in each case is classified into easy migration and limited awareness. Easy migration indicates that the cell migration process occurs too frequently due to the lower thresholds, ψ_{th} and β_{th}. Limited awareness means that the awareness of an eNB about the correct directions of areas with high traffic levels is limited because of the inappropriate angular interval, $\Delta\theta$. The

effects of DOA estimation error on the awareness of eNBs are also analysed.

The QoS performance is shown in Fig. 6.7. Since the UE average transmission power shows an inverse trend with the area power consumption as demonstrated before, only the area power consumption is shown in Fig. 6.8. Due to the exemption of load information exchange delivered by the TACM algorithm, only the switching on-off frequency is taken into account when analysing system overheads in Fig. 6.9. All the parameters for each variation are summarised in Table 6.1.

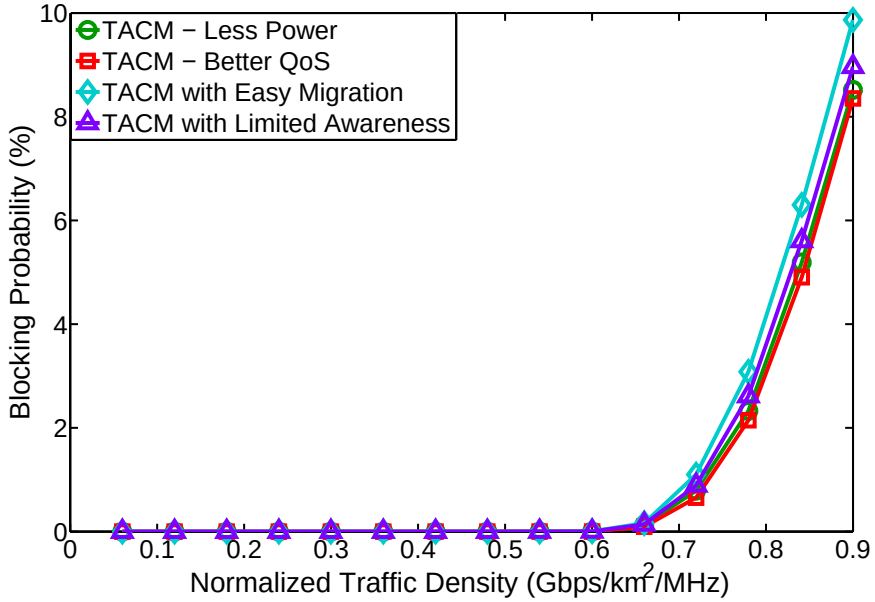
Impact of the Tuning Parameters

The tuning parameters are H_{div}, $DivL_{\text{th}}$, H_{die} and $DieL_{\text{th}}$. H_{div} and $DivL_{\text{th}}$ are designed as a pair to control the cell division while H_{die} and $DieL_{\text{th}}$ control the cell death as a pair. To summarise, bigger H_{div} and $DivL_{\text{th}}$ make cell division less likely to happen, leading to fewer active eNBs in a SCN. Bigger H_{die} and smaller $DieL_{\text{th}}$ make cell death harder to be triggered, leading to fewer sleeping eNBs in a SCN. The tuning parameters in the simulation are varied to first ensure the QoS (blocking probability and delay) is similar to the no sleep mode case. Then they are varied to reduce more power consumption until QoS just starts to degrade. In practical networks, this can be easily done through a software defined architecture with a feedback of the QoS measurements.

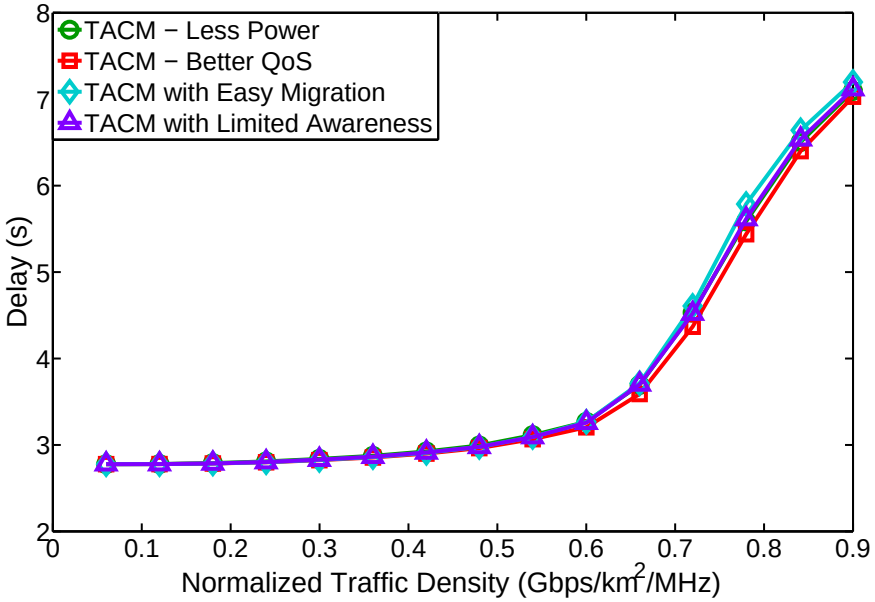
In the better QoS version of TACM with the arguments given in Table 6.1, these pairs of parameters are varied to keep more eNBs active in the network. As shown in Fig. 6.7, the blocking probability and delay is improved a little at the cost of highly increased network power as shown in Fig. 6.8 compared with the less power version of the TACM algorithm. This is achieved by making cell division happen more easily and making cell death a little harder to be triggered. More remarkable is the fact that the QoS improvement gain brought by more active eNBs in the network is limited compared with the larger area power consumption.

Impact of Easy Migration

As illustrated before, ψ_{th} is defined to ensure that cell migration is only triggered based on a sufficiently large amount of DOA records. Like β_{th}, it also controls the frequency of cell migration. When ψ_{th} and β_{th} are set to low values such as the example of easy migration (as shown in Table 6.1) based on the better QoS version, weight bias and memory size may



(a)



(b)

Figure 6.7: Comparisons of the different versions of the TACM algorithm on QoS. (a) Average blocking probability against network traffic density. (b) Average delay against network traffic density.

exceed the thresholds more easily due to small random traffic variations and easy migration occurs then. As shown in Fig. 6.7, easy migration makes the original TACM degrade the QoS and switch more eNBs to sleep modes (as shown in Fig. 6.8). Since there is a trade-off between QoS and network power reduction, it is more convincing to compare the TACM with easy migration to the less power version of TACM, which is configured by varying the tuning parameters without unreliable migrations. Not surprisingly, the reduced power version of

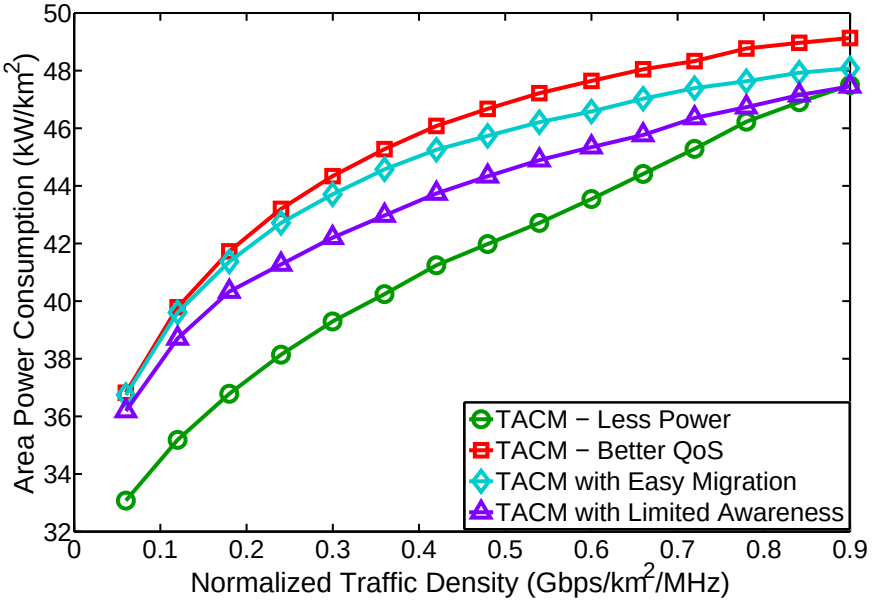


Figure 6.8: Comparisons of the different versions of the TACM algorithm on area power consumption.

the TACM algorithm shows both better QoS and higher network power reduction than the TACM with easy migration. This indicates that the tuning parameters should be used to tune the relationship between QoS and network power reduction if required instead of ψ_{th} and β_{th}. Furthermore, as shown in Fig. 6.9, the redundant migration processes also increase the overall switching on-off times, which add to the system overheads. ψ_{th} should be set big to ensure a big database acquired, where 100 is enough and β_{th} should be relatively high, say 0.1, to reflect a high angular traffic bias. However, the performance of the TACM algorithm is not that sensitive to the variations of these two parameters. Only in the extreme case in the given example, there are notable effects.

Impact of Limited Awareness

The other parameter, $\Delta\theta$, is critical to cell division and cell migration because it determines how to classify the DOAs and how to update the weights, which are decision-making materials of eNBs. A low value of $\Delta\theta$ leads to the situation that an eNB may be too concentrated on the specific small direction ranges without considering the traffic in the angularly adjacent directions. Moreover, an eNB is easier to be enticed by the random traffic spikes from the narrow direction ranges leading to useless cell migration. Comparatively, an over-large $\Delta\theta$ may result in too large a interval $\mathcal{I}_{i,j}$ for $N_{i,j}$, which may not effectively reveal the hotspot directions. This means that an eNB may consider a wide direction range when classifying

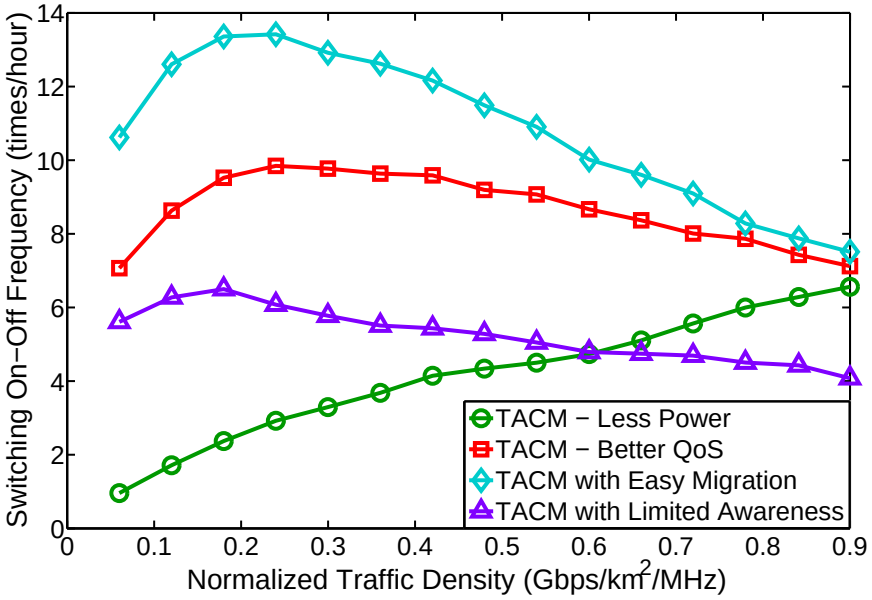


Figure 6.9: Comparisons of the different versions of the TACM algorithm on the switching on-off frequency.

DOAs and update the corresponding weight. This is likely to result in a situation where an eNB regards the direction of a neighbour eNB as the direction of the area with a high traffic level even if the two directions are far apart. These two situations both affect the accuracy of the orientational awareness about the areas of high traffic levels, which are defined as limited awareness. An example of the TACM with limited awareness is given by setting a small $\Delta\theta$ based on the better QoS version of TACM and its arguments are listed in Table 6.1. Similarly, comparing it with the less power version, it shows slightly worse QoS but significantly higher area power consumption as shown in Fig. 6.7 and Fig. 6.8, respectively. This reveals the importance of choosing a medium and appropriate $\Delta\theta$. However, there is a relatively big range for choosing $\Delta\theta$ and the performance only begins to decline in extreme cases. Due to this feature, the TACM algorithm is actually very strong in defending inappropriate DOA classifications.

Impact of the DOA Estimation Error

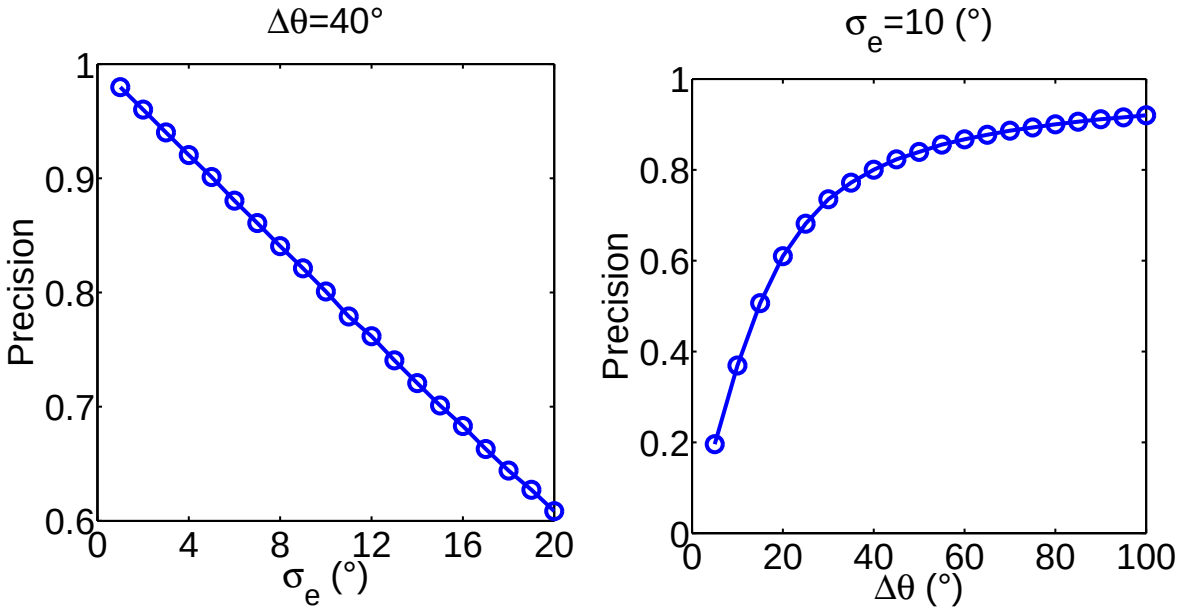
In some applications in the past, the accuracy of DOA estimation has proved problematic [89, 90]. However, as mentioned before, the probability of having NLOS is low in the investigated SCN scenario and the accuracy of DOA estimation can be very high. Moreover, the approach of using DOAs in the TACM algorithm is somewhat different as a relatively coarse estimation is required to identify the sectors which UEs are located in. From the

Table 6.1: Summary of the TACM algorithm parameters

TACM Parameters	TACM - Less Power	TACM - Better QoS	TACM with Easy Migration	TACM with Limited Awareness
H_{div}	5s	4s	4s	4s
$DivL_{\mathrm{th}}$	8%	4%	4%	4%
H_{die}	40s	50s	50s	50s
$DieL_{\mathrm{th}}$	0%	0%	0%	0%
$\Delta\theta$	40°	40°	40°	4°
ψ_{th}	100	100	10	100
β_{th}	0.1	0.1	0.01	0.1

aforementioned performance discussion, the TACM algorithm is shown to be fault-tolerant in terms of DOA classification. To further check this assumption, the impact of DOA estimation error is briefly investigated. A normal distribution with zero mean is adopted to model the DOA estimation error as an example. It is applied on the DOAs which are intended to be classified into sectors with an angular interval $\Delta\theta$ (explained in section 2.2). Precision and recall are utilised to evaluate the effect of the DOA estimation error. Precision is defined as the number of 'true positives' (i.e. the number of DOAs correctly classified into the interval) divided by the total number of DOAs classified into the interval. Recall is defined as the number of 'true positives' divided by the total number of DOAs that actually ought to be classified into the interval. In the case investigated, recall has the same tendency as precision so that only precision is shown in Fig. 6.10 (a) and (b).

Fig. 6.10 (a) gives an example of classification precision against the variations of the standard deviation of the normal distribution σ_{e} when setting the angular interval $\Delta\theta$ to 40°. From the figure, it is shown that the probability of having a wrong DOA classification is low when the DOA estimation error is low for a reasonable angular interval (40°), e.g. 80% precision corresponds to a standard deviation of 10°. In Fig. 6.10 (b), classification precision is plotted against different angular intervals revealing that precision of DOA classification is always high if the angular interval is chosen to be medium or high. With reasonable angular intervals chosen in the TACM algorithm, the effect of DOA estimation errors is not significant, meaning that an ideal DOA estimation assumption is sufficient. This can be confirmed by setting $\Delta\theta$ to an appropriate value to avoid the low precision. If σ_{e} is 10°, with $\Delta\theta$ set to 40°, almost 99.7% of DOAs have absolute errors smaller than 30 degrees.



(a) (b)

Figure 6.10: An investigation of DOA estimation precision. (a) Precision of DOA classification affected by DOA estimation error with fixed classification angular interval against the standard deviation of the error distribution. (b) Precision of DOA classification affected by DOA estimation error with fixed error standard deviation against the variations of the angular interval.

With the error applied in the simulation, the resulting DOA estimation error does not affect the performance of the original TACM algorithm, but it starts to affect the overall performance of its less power version, where the system overhead produced by switching on-off is especially exacerbated relatively. However, generally speaking, as the effect of DOA estimation error in the performance of the TACM algorithm is limited even when σ_e is not small, the TACM proves robust against the DOA estimation inaccuracy if $\Delta\theta$ is properly chosen. Furthermore, DOA estimation is not used for positioning UEs in this case, but is used to reflect the relative arriving angles of signals, which better reveal the radio environments.

6.6 Conclusion

A novel sleep mode operation algorithm for distributed SCNs is proposed in this chapter to overcome the drawbacks of the ATP based sleep mode operation proposed in Chapter 5. The algorithm employs an innovative concept, TACM, to achieve the objectives of no computation at sleeping eNBs, fewer system overheads and less UE transmission power. In the TACM, cell activities are controlled, either cell division, cell death or cell migration, which

accompany state transitions of eNBs. The TACM algorithm enables each eNB to be aware of the directions of hotspot areas with the application of DOAs as traffic information, which completely remove the necessity of load information exchange for sleep mode operation required by other traditional sleep mode operation schemes. The simulation results based on SCNs with non-ideal distributions show that the TACM algorithm achieves more than 34% reduction in area power consumption when the traffic level is low, which is equivalent to 86% less overall active time compared with the system without sleep mode operation. Compared with a consolidated baseline algorithm, the TACM algorithm significantly mitigates the system overheads and notably reduces the increase of UE transmission power. Assuming high energy consumption of transient states, the TACM algorithm can reduce the area power consumption by up to 81% compared with the baseline scheme. The TACM algorithm is also shown to be insensitive to the direction of arrival estimation inaccuracy.

Chapter 7

Hotspot-oriented Green Frameworks

Contents

7.1 Introduction

The sleep mode operation algorithms proposed in the previous two chapters are for distributed SCNs, such as the current LTE system. As indicated by the hypothesis of the thesis, sleep mode operation should be considered for different architectures. In terms of the architecture for SCNs with centralised management entities, Cloud RAN (C-RAN) is a strong candidate for future SCNs. It has improved energy efficiency as one of the advantages aforementioned in Chapter 2. In a C-RAN, remote radio heads (RRHs) are adopted to perform radio functions,

providing signal coverage. A Baseband Unit (BBU) pool is placed at a centralised site, where BBUs are aggregated and deployed at multiple general purpose servers [15, 16, 18, 76]. Centralisation of the BBUs provides the advantage that some servers can be switched to sleep modes for energy saving when the overall demand of processing resources is low [3,48,91]. Just like base stations in conventional architectures, RRHs can also be switched to low power sleep modes due to temporal and spatial traffic variations in a C-RAN. However, the difference in the architecture requires different sleep mode operation schemes, which can optimise the state transitions from a global view of a C-RAN at the central site to achieve its centralisation gain.

At the same time of changing the locations of controlling engines, the objectives of sleep mode operation remain the same as introduced in the last chapter, including maintaining QoS while reducing the network power consumption, no computation at sleeping RRH and fewer system overheads. The increase in UE transmission power should also be minimised by placing the active RRHs nearer to the hotspot areas. The placement of active RRHs should be optimised with the collected traffic information from RRHs to exploit the centralisation advantage.

In this chapter, different availabilities of traffic information are also studied considering various infrastructure conditions. Hotspot-oriented Green Frameworks are proposed, in which different clustering strategies are adopted at central controllers, managing sleep mode operation from a global view and placing active RRHs closer to hotspot areas. Each of the frameworks is dedicated to a case with a certain completeness in traffic information and the impact of different availabilities is also investigated.

In Section 7.2, the problem of hotspot areas in C-RANs is analysed with a benchmark given. The analysis directly adopts a result proposed in Chapter 4. Section 7.3 proposes the details of the Hotspot-oriented Green Frameworks as the solutions to reducing the UE transmission power for C-RANs with different infrastructure availabilities. In Section 7.4, the performance of the Hotspot-oriented Green Frameworks is evaluated and compared with a baseline strategy. The conclusions are drawn in Section 7.5.

7.2 Problem Analysis and Solution

Temporal and spatial traffic variations in a C-RAN create an unbalanced traffic distribution and form hotspot areas with relatively high traffic levels. This leaves space for energy saving

where the local traffic levels are low, but also brings a major challenge that is to provide just enough radio resources from a global view and supply the resources according to the local demands. The benefits in terms of power consumption are the reduction in the total network power consumption and the discounted increase in the average UE power consumption, respectively, which are two of the objectives of sleep mode operation in a C-RAN. In this section, the problem caused by hotspot areas is analysed based on the stochastic MHPP model introduced in Chapter 3. The analysis points out the key point of sleep mode operation and provides a benchmark for the average UE transmission power by integrating the results proposed in Chapter 4. The solution to the problem, the clustering technique, is introduced.

7.2.1 Problem of Hotspot Areas

Considering a C-RAN implemented with a two-state sleep mode operation, the RRHs are represented as $\mathcal{B} = \{B_b\} = \mathcal{B}^{\text{on}} \cup \mathcal{B}^{\text{off}}$ $(b \in \mathbb{N}^*, b \leq |\mathcal{B}|)$, where active RRHs are denoted as $\mathcal{B}^{\text{on}}$ and $\mathcal{B}^{\text{off}}$ are RRHs at sleep modes. For a C-RAN modelled by a MHPP Φ_{m} of intensity λ_{m}, the derivative of $F(r)$ (proposed in Section 4.4.1), $f(r)$, which is the probability density function of the distance between u and its nearest RRH, can capture the distance between an UE of an arbitrary location and a RRH in a C-RAN, and can be used for calculating the average UE transmission power. Sleep mode operation in the C-RAN can be treated as a second thinning process based on $\Phi_{\text{m}} \cap \mathcal{A}$. The corresponding retaining probability $\eta_t(A)$ at time t depends on the sleep mode operation strategy adopted then and the exact location $A \in \mathcal{A}$. Naturally, it relates to the proportion of the active RRHs and thus area power consumption of the C-RAN $\overline{P_t^{\text{N}}}$ contributed by the RRHs is

$$\overline{P_t^{\text{N}}} = \int_{\mathcal{A}} \frac{\eta_t(A)\lambda_{\text{m}}P_{\text{B}} + \big(1 - \eta_t(A)\big)\lambda_{\text{m}}P_{\text{S}}}{|\mathcal{A}|}\mathrm{d}A, \tag{7.1}$$

where P_{B} and P_{S} are the RRH power consumption at the active mode and the sleep mode, respectively. As $P_{\text{S}} < P_{\text{B}}$, $\overline{P_t^{\text{N}}}$ can be reduced with lower $\eta_t(A)$. With the sleep mode operation, the resulting $f(r)$ can be adjusted to $f_t\big(r, \eta_t(A)\big)$, where $\forall A \in \mathcal{A}, \int_0^\infty f_t\big(r, \eta_t(A)\big)\mathrm{d}r = 1$ should be satisfied.

A generic UE $u \in \mathcal{A}$ in a typical C-RAN system implemented with the sleep mode operation described by $\eta_t(A)$ can be assumed to be served by its nearest active RRH with open-loop power control for uplink, of which the required signal-to-noise ratio (SNR) is β. For an arbitrary pattern of the active UE distribution knowing the location-dependent intensity

$\lambda_{\mathrm{u}}(A)$, the average uplink UE transmission power $\overline{P_t^{\mathrm{u}}}$ can therefore be calculated as

$$\overline{P_t^{\mathrm{u}}} = \frac{\int_{\mathcal{A}} \lambda_{\mathrm{u}}(A) \int_0^{\infty} \beta\sigma^2\alpha(r) f_t\Big(r,\eta_t(A)\Big)\mathrm{d}r\mathrm{d}A}{\int_{\mathcal{A}} \lambda_{\mathrm{u}}(A)\mathrm{d}A}, \tag{7.2}$$

where σ^2 is the noise power and $\alpha(r)$ is the attenuation of the propagation path between u and its connecting RRH. For lower $\eta_t(A)$ to reduce $\overline{P_t^{\mathrm{u}}}$, $f_t\Big(r,\eta_t(A)\Big)$ is skewed to larger r with a longer tail considering the curve shape shown in Fig. 4.5 (e) in Section 4.4.4 and yields higher expected UE transmission power $P_t^{\mathrm{u}}(A)$ at A by a bigger integral with respect to r. Thus, $\overline{P_t^{\mathrm{u}}}$ is raised owing to the reduced number of active RRHs.

With the application of sleep modes, the first operation target is to ensure that $\forall A \in \mathcal{A}$ the QoS (e.g. blocking probability, delay) at A is not degraded when $\eta_t(A) \neq 1$ compared with the QoS when $\eta_t(A) = 1$. On this constraint, $\overline{P_t^{\mathrm{N}}}$ given by Equation (7.1) should be maximised. From a global view of a centralised controller, it has to first determine $\overline{\eta_t} = \int_{\mathcal{A}} \eta_t(A)/|\mathcal{A}|\,\mathrm{d}A$, which is the basic objective but usually difficult to solve without knowing the overall QoS when no sleep mode operation is applied.

The problem is more complicated by considering the heterogeneous local service demands that are likely to be caused by a highly unbalanced $\lambda_{\mathrm{u}}(A)$ due to the existence of hotspot areas, $\eta_t(A)$ should accommodate various demands for different A and t even for a fixed $\overline{\eta_t}$. In terms of the average UE uplink transmission power for a given $\overline{\eta_t} < 1$, the increase in $\overline{P_t^{\mathrm{u}}}$ due to sleep mode operation can be reduced with lower $P_t^{\mathrm{u}}(A)$ for A, where $\lambda_{\mathrm{u}}(A)$ is large, by having $f_t\Big(r,\eta_t(A)\Big)$ skewed to small r, i.e. with higher $\eta_t(A)$. This means that active RRHs should be put close to hotspot areas.

Owing to various UE distribution patterns, $\overline{P_t^{\mathrm{u}}}$ in a C-RAN cannot be determined if $\lambda_{\mathrm{u}}(A)$ is not accurately modelled. However, considering an arbitrary $\lambda_{\mathrm{u}}(A)$ distribution, the benchmark average UE transmission power $\widetilde{P_t^{\mathrm{u}}}$ can be obtained by replacing $\eta_t(A)$ in Equation (7.2) with $\overline{\eta_t}$ as

$$\widetilde{P_t^{\mathrm{u}}} = \int_0^{\infty} \beta\sigma^2\alpha(r) f_t(r,\overline{\eta_t})\mathrm{d}r. \tag{7.3}$$

It assumes a uniform selection of active RRHs in sleep mode operation, giving equal $P_t^{\mathrm{u}}(A)$ for all A. Taking it as the benchmark, an effective sleep mode operation strategy should make $\eta_t(A)$ accommodative to hotspot areas and achieve $\overline{P_t^{\mathrm{u}}}$ lower than $\widetilde{P_t^{\mathrm{u}}}$.

7.2.2 Solution

As analysed, sleep mode operation is required to place active RRHs at the appropriate hotspot places, reducing the average distance between active RRHs and UEs. Clustering techniques can be utilised to achieve the on-demand provisioning of radio resources, ensuring local QoS requirements and minimising the increased average UE transmission power. In wireless communications, clustering can be used to group network components or other network resources based on their radio environments or parameters for further utilisation. Clustering techniques are widely used in ad-hoc networks to guarantee basic performance achievements [92, 93]. Some work has also been done to adapt the conventional clustering techniques from the ad-hoc network domain to the cellular network domain. In [94–96], clustering is applied to femtocell networks, where femtocell base stations are grouped into clusters and further intra-cluster resource management can then be implemented. The clustering techniques have also been applied to distributed architectures for sleep mode operation. In [97–99], base stations in a network are formed into clusters so that some formats of information can be exchanged within clusters for sleep mode control. However, there is still no work done to cluster a quantified traffic distribution for sleep mode operation in a fully centralised architecture such as C-RAN, which is considered in the following context.

Clustering is a technique for partitioning a set of d-dimensional data entries, $\mathcal{X} = \{x_i\}$, into k ($\leq |\mathcal{X}|$) subsets or clusters, $\mathcal{S}_j$ ($i, j \in \mathbb{N}^*, i, j \leq k$). Data entries are some formats of information under consideration and the objective of the clustering process is to find the similarities among such data entries. Then data entries are grouped based on the similarities, and cluster centroids are found to represent them.

Thanks to the above function, clustering can be used to partition the quantified traffic distribution pattern in a C-RAN. The traffic having similarity in geographical location can be represented by a cluster centroid, meaning that the data entry set $\mathcal{X}$ should store location information of the traffic. Then in sleep mode operation, each cluster centroid can be used to determine the location of one closely located active RRH, expected to serve the quantified traffic at the corresponding local area. Therefore, sleep mode operation benefits from the final objective of the clustering process that is to find the proper cluster centroids in the format of locations rather than grouping data entries into clusters. By determining the locations of all active RRHs in this way, a minimised average distance between UEs and RRHs can be achieved and the average UE transmission power can be minimised as well.

As a widely used clustering algorithm, K-means can be adopted and customised to suit a C-RAN architecture for sleep mode operation. In the considered scenario with data representing locations, the purpose of the clustering process is to find the cluster centroids, where the sum of all the within-cluster distances between the cluster members and their respective cluster centroids are minimum. Therefore, by considering the distances as the Euclidean distances, the objective of a clustering process is to minimise the cost function defined as

$$J(\mathcal{S}_1,\ldots,\mathcal{S}_j,\ldots,\mathcal{S}_k,C) = \sum_{j=1}^{k} \sum_{x_i \in \mathcal{S}_j} \|x_i - c_j\|^2, \tag{7.4}$$

where $C = \{c_j\}$ and c_j is the cluster centroid of the cluster S_j. $\|x_i - c_j\|$ is the Euclidean distance between x_i and c_j.

At the start of the clustering process, a pre-determined number of (k) locations are chosen as the initial cluster centroids $c_j^{(0)}$, where the superscript denotes the iteration number, l ($l \in \mathbb{N}^*$). In iteration 1, each location data entry, x_i, is associated with one $c_j^{(0)}$, which has the smallest distance to x_i. All x_i associated with the same $c_j^{(0)}$ constitute a cluster $S_j^{(1)}$ and become its cluster members. After the cluster formation, the mean (of each dimension) of $x_i \in S_j^{(1)}$ is computed and each $c_j^{(0)}$ is replaced by the calculated mean, becoming $c_j^{(1)}$. At this point, iteration 1 is finished and each $c_j^{(1)}$ becomes the target centroid for association in the next iteration. This process starting from centroid association is repeated for a certain number of iterations or until the locations of the cluster centroids do not have great changes. In successive iterations, the cluster centroids are updated, moving to the locations better representing their corresponding cluster members, and the cost function, $J(\mathcal{S}_1,\ldots,\mathcal{S}_j,\ldots,\mathcal{S}_k,C)$, decreases until a stable level.

Being a promising solution, the application of the clustering technique for sleep mode operation in C-RANs has unsolved issues, which are solved in the proposed Hotspot-oriented Green Frameworks introduced in the next section.

7.3 Hotspot-oriented Green Frameworks

As aforementioned, the basic idea of adapting the clustering technique to sleep mode operation is mapping an active RRH to a cluster centroid, which represents the location of the local traffic. From a global view, the result is that the active RRHs are placed where the traffic is, i.e. closer to hotspot areas. Taking it as the basis, the number of the active RRHs in a C-RAN

is equal to the number of the cluster centroids of a clustering process. The Hotspot-oriented Green Frameworks proposed in this section are designed to involve the application of the clustering technique for sleep mode operation dedicated to C-RANs of different information availabilities. The frameworks solve some critical issues related to transplanting the clustering technique, which are presented as follows.

- *Quantifying the local traffic based on information availabilities:* The traffic quantification is a common point of all sleep mode operation schemes threaded by the hypothesis of the thesis. With the application of the clustering technique, the local traffic relative to each RRH in a C-RAN should be quantified at discrete locations as data entries of the clustering process. Practical information and infrastructure availabilities should be considered for different C-RAN architectures because the performance of the clustering process depends on the format of the traffic information provided. Generally speaking, the traffic can be better captured with more location information available. The information and infrastructures needed in different environments are considered in the proposed frameworks.

- *Choosing an appropriate clustering frequency:* Clustering frequency refers to the frequency of running the clustering process at the centralised controller to adjust the RRH on-off states in a C-RAN. It also refers to the frequency of evaluating network conditions, e.g. variations of hotspot areas and changes of network load levels, from past experiences. A high cluster frequency usually makes a C-RAN respond to traffic variations more swiftly, while a low cluster frequency requires fewer computing resources and yields less RRH switching on-off overhead. The choices of the clustering frequency and the corresponding trade-offs are presented with the analysis of the simulation results.

- *Determining an appropriate number of cluster centroids:* As the number of cluster centroids in a clustering process is equal to the number of active RRHs in a C-RAN, it needs to be determined properly to reflect the overall traffic level so that just enough radio resources are provided, ensuring a similar QoS and maximum power consumption can be reduced. It can be determined by evaluating the average network load as mentioned later.

- *Determining appropriate initial cluster centroids:* The initial cluster centroids, $c_j^{(0)}$, should be appropriately pre-determined to start clustering process. Through clustering iterations in a clustering process, they affect the final choice of cluster centroids and

therefore the locations of the active RRHs. An extensive adjustment of RRH states at one time instance may result in network instability and yield large system overheads. It can be avoided by initialising the clustering centroids based on the locations of the current active RRHs with some adjustments. More details will be given in the later explanation.

- *Adjusting cluster centroids to real RRH locations:* Since RRH locations in a C-RAN are spatially discrete while computed cluster centroids are continuous locations, the cluster centroids should be moved to real RRH locations at every iteration during a clustering process so that the clustering process can minimise the cost function based on real RRH locations. Furthermore, the case where multiple cluster centroids are adjusted to one real RRH location should be avoided in order to maintain the number of active RRHs equal to the wanted pre-determined number. To address this, virtual cluster centroids are introduced and are associated with real RRH locations during a clustering process. This process will be described in detail later.

7.3.1 Complete Framework

In the *Complete Framework*, the aforementioned local traffic distribution is depicted and quantified based on the UE location information. Due to the spread of Location Based Services, UE location information is much easier to obtain and therefore can be utilised. The location of a UE can be estimated by an appropriate satellite navigation system in outdoor scenarios or other positioning technologies, such as those using Direction of Arrival and Time of Arrival information, in indoor scenarios.

UE location information can be reported and collected at the central controller to perform the clustering process because of the centralisation of a C-RAN. In the case where each UE has its location information available when requesting services, data entries $x_i \in \mathcal{X}$ of a clustering process are just the locations of the service requests, which are effectively a part of the traffic information. The clustering process then partitions the locations into clusters and the locations correspondingly become the cluster members. The location information of the RRHs, $\mathcal{H} = \{h_b\}$, is acquired through manual measurements by operators and recorded in the controller when deployed. In a self-organised network, they can also be self-measured during self-configuration and may be self-optimised through a long-term measurement.

The clustering frequency is defined as $1/T$, where T is the clustering period and the central

controller runs the clustering algorithm every T seconds. During the clustering period, the location information of UEs is collected by the RRHs delivering services and is reported to the centralised controller when UEs request services. Every reported location is treated as a data entry, x_i, for the next clustering process. Meanwhile, the loads of all active RRHs are sent to and recorded at the central controller as the other part of the traffic information whenever there is a load variation at a RRH. At the end of a clustering period, the average overall load $\overline{L}$ in the last T seconds is estimated as

$$\overline{L} = \sum_{B_b \in \mathcal{B}^{\text{on}}} \int_0^T \frac{L_{b,t}}{|\mathcal{B}^{\text{on}}|\,T} \mathrm{d}t, \tag{7.5}$$

where $L_{b,t}$ is the load of B_b at time t.

The number of the cluster centroids of the next clustering process, k, i.e. the number of the active RRHs on completion of the next clustering process, depends on $\overline{L}$, the number of the current active RRHs in the C-RAN, $|\mathcal{B}^{\text{on}}|$, and a RRH load reference, L_{ref}, based on the principle that

$$k \leftarrow \min(\lceil \frac{\overline{L}\,|\mathcal{B}^{\text{on}}|}{L_{\text{ref}}} \rceil, |\mathcal{B}|). \tag{7.6}$$

The principle yields a larger k when $(\overline{L} > L_{\text{ref}}) \wedge (|\mathcal{B}^{\text{on}}| < |\mathcal{B}|)$ and a smaller k when $(\overline{L} < \frac{|\mathcal{B}^{\text{on}}|-1}{|\mathcal{B}^{\text{on}}|} L_{\text{ref}}) \wedge (|\mathcal{B}^{\text{on}}| > 1)$. In other cases, k remains the same. This is equivalent to the fact that more RRHs will be active when the overall load of the RRHs is above a threshold and some RRHs will be switched to sleep modes when the overall load of the RRHs is below a threshold. If the calculated average load of the clustering period is between these two thresholds, the number of the active RRHs does not change. No matter which case it is, the locations of the active RRHs after a clustering process may change based on the data entries provided.

When the clustering starts with the pre-determined k clusters (cluster centroids), the locations of the current $\mathcal{B}^{\text{on}}$ are considered as the candidate cluster centroids. If the number of the active RRHs are intended to increase, $k - |\mathcal{B}^{\text{on}}|$ more cluster centroids, which are the locations of $\mathcal{B}^{\text{off}}_{k-|\mathcal{B}^{\text{on}}|}$, are randomly chosen from the locations of the current $\mathcal{B}^{\text{off}}$. Together with the candidate cluster centroids, they form the initial cluster centroids $\mathcal{C}^{(0)}$. On the contrary, if some RRHs are to be deactivated, $|\mathcal{B}^{\text{on}}| - k$ cluster centroids are randomly chosen and removed from the candidate cluster centroids. The rest of them, the locations of $\mathcal{B}^{\text{on}}_k$ (as a subset of $\mathcal{B}^{\text{on}}$), form $\mathcal{C}^{(0)}$. If the number of the active RRHs remains the same, the candidate centroids are selected as $\mathcal{C}^{(0)}$ for the cluster centroid updates through the iterations

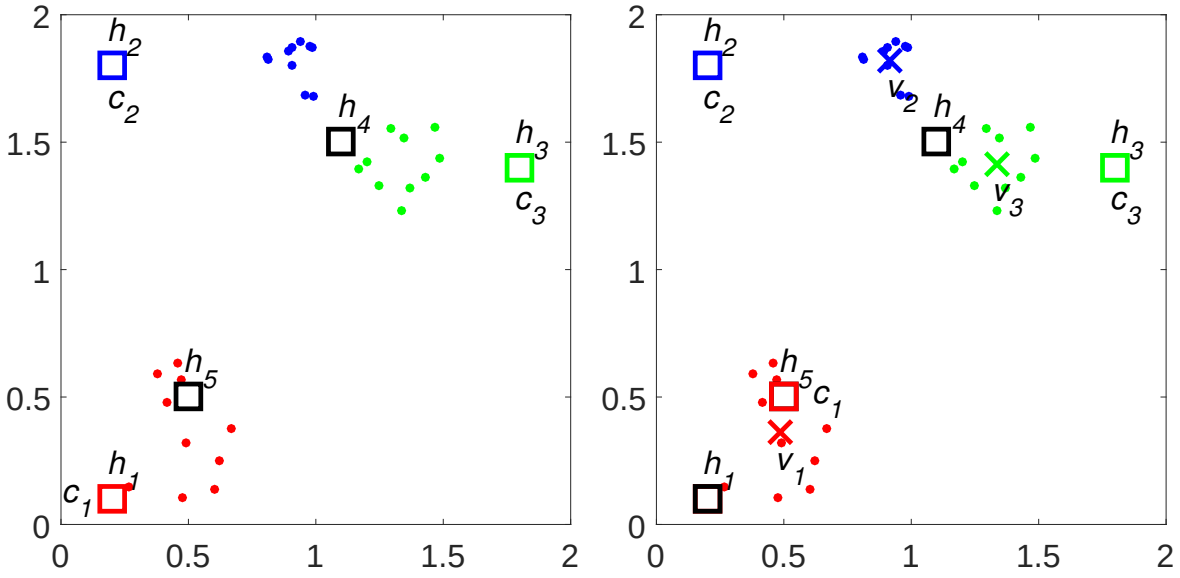
in a clustering process. Thus, the locations of the active RRHs on completion of the clustering process are based on the locations of the active RRHs before the clustering process to avoid a significant adjustment of the RRH on-off states and extra system overheads.

As RRHs in a C-RAN are given discrete spatial locations $\mathcal{H} = \{h_b\}$ when deployed, the updated cluster centroids of continuous locations should be linked to $\mathcal{H}$ of the C-RAN. To ensure that the clustering process is minimising the cost function, $J(\mathcal{S}_1,\ldots,\mathcal{S}_j,\ldots,\mathcal{S}_k,\mathcal{C})$, based on the actual C-RAN layout, h_b should be taken into consideration for updating the cluster centroids c_j in every clustering iteration l. Thus, instead of directly replacing $c_j^{(l-1)}$ in iteration l with the computed within-cluster means of all cluster members x_i in a cluster $\mathcal{S}_j^{(l)}$, they are treated as the virtual cluster centroids, $v_j^{(l)} \in \mathcal{V}^{(l)}$ as

$$v_j^{(l)} = \frac{\sum_{x_i \in \mathcal{S}_j^{(l)}} x_i}{\left|\mathcal{S}_j^{(l)}\right|}. \tag{7.7}$$

For each $v_j^{(l)}$ of $\mathcal{S}_j^{(l)}$, the closest (in terms of the Euclidean distance) h_b is assigned as $c_j^{(l)}$ of iteration l. If there are multiple $v_j^{(l)}$ sharing the same h_b, $c_j^{(l-1)}$ are not updated to $c_j^{(l)}$ in iteration l to avoid merging multiple cluster centroids into one. Otherwise, there will be fewer active RRHs than the pre-determined number of cluster centroids. In the example shown in Fig. 7.1 (a), the clusters start with h_1, h_2 and h_3 (shown as squares) as the cluster centroids, which are c_1, c_2 and c_3, respectively. Then, the means of the cluster members (the points) in all clusters are computed and defined as the virtual cluster centroids, v_1, v_2 and v_3, shown as crosses in Fig. 7.1 (b). The closest RRH location relative to each virtual cluster centroid is found, which is h_5 for v_1, h_4 for v_2 and v_3. The cluster centroid c_1 is updated to h_5 while c_2 and c_3 do not change because the corresponding virtual centroids are intended to move to the same location, h_4. By updating cluster centroids in such method, the association of data entries to the clusters is always based on real RRH locations from the first iteration.

The cost function, $J^{(l)}(\mathcal{S}_1^{(l)},\ldots,\mathcal{S}_j^{(l)},\ldots,\mathcal{S}_k^{(l)},\mathcal{C}^{(l)})$, is calculated before the end of iteration l in a clustering process and compared with the cost function of iteration $l-1$, $J^{(l-1)}$. The clustering iteration may cease at this point if the difference between two cost functions, $|J^{(l)} - J^{(l-1)}|$, is smaller than 1% of $J^{(l-1)}$, or if $l = 10$. This is to guarantee that the clustering process has reached the state where the average distance between data entries and their cluster centroids are reduced to a relatively stable level. Through simulation, it is observed that more iterations do not significantly decrease the final values of the cost function.



(a) (b)

Figure 7.1: Example of a clustering process of points as data entries. (a) Before the update of cluster centroids. (b) After the update of cluster centroids.

After the clustering process, deactivation signals are sent to the active RRHs if the final cluster centroid set does not contain their locations, and then the corresponding active RRHs are switched to sleep modes. Activation signals are transferred to the sleeping RRHs if their locations belong to the final cluster centroid set, and the corresponding sleeping RRHs are switched to active modes. The data entry memory is cleared for the next clustering period and the timer is set to zero. For the active RRHs to be deactivated, it stops admitting UEs and remains active until the data transmissions of all the associated UEs are finished. After state transitions, all the active RRHs expand or shrink their cell sizes to ensure coverage and the UEs re-associate to the closest active RRHs in terms of path loss.

Through such cycles, the centralised controller exploits $x_i \in \mathcal{X}$ and $L_{b,t}$ as the traffic information for the dynamical adaptation of the RRH states to the overall RRH loads and variations of the hotspot areas. The clustering process is run based on a massive amount of data collected, resulting in a highly accurate inference of the hotspot areas with a large number of service requests in a clustering period. The active RRHs can always be placed where needed because of the complete information collected, reducing the average distance between UEs and RRHs, which can be regarded as a upper bound of the reduction in the increased average UE transmission power. The pseudo-code of *Complete Framework* is given in **Algorithm 1**.

Algorithm 1 Complete Framework

for all $B_b \in \mathcal{B}^{\text{on}}$ **do**
  Report the load variations to the central controller
  Report the locations of received service requests to the central controller
  if Receive deactivation signal from the controller **then**
    Switch to the sleep modes
  end if
end for
for all $B_b \in \mathcal{B}^{\text{off}}$ **do**
  Do nothing
  if Receive activation signal from the controller **then**
    Switch to the active modes
  end if
end for
for the central controller **do**
  Run the timer
  Record the RRH load variations
  if T seconds has elapsed **then**
    Record the locations of service requests as x_i, $\mathcal{X} \leftarrow \mathcal{X} \cup \{x_i\}$
    $\overline{L} \leftarrow \sum_{B_b \in \mathcal{B}^{\text{on}}} \int_0^T \frac{L_{b,t}}{|\mathcal{B}^{\text{on}}|T} \mathrm{d}t$
    $k \leftarrow \min(\lceil \frac{\overline{L}|\mathcal{B}^{\text{on}}|}{L_{\text{ref}}} \rceil, |\mathcal{B}|)$
    $l \leftarrow 0$
    if $k > |\mathcal{B}^{\text{on}}|$ **then**
      $C^{(l)} \leftarrow \{h_b \in \mathcal{H} : B_b \in (\mathcal{B}^{\text{on}} \cup \mathcal{B}^{\text{off}}_{k-|\mathcal{B}^{\text{on}}|})\}$
    else if $k < |\mathcal{B}^{\text{on}}|$ **then**
      $C^{(l)} \leftarrow \{h_b \in \mathcal{H} : B_b \in \mathcal{B}^{\text{on}}_k\}$
    else
      $C^{(l)} \leftarrow \{h_b \in \mathcal{H} : B_b \in \mathcal{B}^{\text{on}}\}$
    end if
    $J^{(l)} \leftarrow \infty$
    Do nothing
    repeat
      $l \leftarrow l + 1$

33: $\forall x_i \in \mathcal{X}, x_i \mapsto \arg\min_{c_j^{(l-1)} \in \mathcal{C}^{(l-1)}} \|x_i - c_j^{(l-1)}\|^2$

34: $\mathcal{S}_j^{(l)} \leftarrow \{x_i \in \mathcal{X} : x_i \mapsto c_j^{(l-1)}\}$

35: $v_j^{(l)} = \frac{\sum_{x_i \in \mathcal{S}_j^{(l)}} x_i}{|\mathcal{S}_j^{(l)}|}$

36: $c_j^{(l)} \leftarrow \arg\min_{h_b \in \mathcal{H}} \|v_j^{(l)} - h_b\|^2$

37: $\mathcal{C}^{(l)} \leftarrow \{c_j^{(l)}\}$

38: $\forall c_j^{(l)} \in (\mathcal{C}^{(l)} - c_j^{(l)}), c_j^{(l)} \leftarrow c_j^{(l-1)}$

39: $J^{(l)}(\mathcal{S}_1^{(l)}, \ldots, \mathcal{S}_j^{(l)}, \ldots, \mathcal{S}_k^{(l)}, \mathcal{C}^{(l)})$
$\leftarrow \sum_{j=1}^{k} \sum_{x_i \in \mathcal{S}_j^{(l)}} \|x_i - c_j^{(l)}\|^2$

40: **until** $\frac{|J^{(l)} - J^{(l-1)}|}{J^{(l-1)}} < 1\% \vee l \geq 10$

41: Send activation signals to $\{B_b \in \mathcal{B}^{\text{off}} : h_b \in \mathcal{C}^{(l)}\}$

42: Send deactivation signals to $\{B_b \in \mathcal{B}^{\text{on}} : h_b \notin \mathcal{C}^{(l)}\}$

43: $\mathcal{X} \leftarrow \varnothing$

44: Set the timer to 0

45: **end if**

46: **end for**

7.3.2 Load Weighted Framework

In some scenarios such as indoor environments, the complete availability of accurate UE locations does not always exist considering the current technologies and the protocols. In these cases, a C-RAN can tolerate the inaccuracy or the incompleteness in UE location information. Otherwise, leaving *Complete Framework* as a strong future candidate, a weighting strategy is considered in *Load Weighted Framework* under the circumstance where only the locations of RRHs in the C-RAN are available.

Although sleep mode operation benefits from the proper locations of cluster centroids rather than grouping UEs into clusters, the loss of UE location information affects the local traffic quantification, which should be realised with other available formats of information to capture the traffic variations. In *Load Weighted Framework*, $\mathcal{X}$ are formed by the locations of the active RRHs in the current clustering period as $\{h_b \in \mathcal{H} : B_b \in \mathcal{B}^{\text{on}}\}$. Since the average load of a RRH in a period of time $\overline{L_b}$ indicates the number of service requests received by the RRH, it can be used to represent its local traffic level during that time. By weighting the locations of $\mathcal{B}^{\text{on}}$ with their respective average load levels, the distribution of the traffic in $\mathcal{A}$ served by $\mathcal{B}^{\text{on}}$ can be roughly quantified. Taking this into consideration at the beginning

of a clustering process, each $x_i \in \{h_b \in \mathcal{H} : B_b \in \mathcal{B}^{\text{on}}\}$ is assigned a weight, $w_i \in \mathcal{W}$. The weight value is equal to $\overline{L_b}$ of B_b, which is defined as

$$\overline{L_b} = \int_0^T \frac{L_{b,t}}{T} \mathrm{d}t. \tag{7.8}$$

Representing the traffic levels at areas served by $\mathcal{B}^{\text{on}}$, $\mathcal{W}$ fundamentally biases a clustering process, moving cluster centroids closer to the areas assigned with higher weights. This is achieved by computing a virtual cluster centroid $v_j^{(l)}$ in iteration l considering $\mathcal{W}$ after the formation of the cluster $\mathcal{S}_j^{(l)}$ as

$$v_j^{(l)} = \frac{\sum_{x_i \in \mathcal{S}_j^{(l)}} x_i w_i}{\sum_{w_i \in \{w_i : x_i \in \mathcal{S}_j^{(l)}\}} w_i}. \tag{7.9}$$

In iteration l, all $c_j^{(l)}$ are chosen based on the biased $v_j^{(l)}$, increasing the probability of placing active RRHs at the areas with high traffic levels and reducing the average UE transmission power. It is worth mentioning that when the number of active RRHs is to increase after a clustering process, the iterations are skipped after the generation of $C^{(0)}$ due to more clustering centroids than the data entries. However, the location optimisation can be processed at the ends of the following clustering periods as long as $k \leq |\mathcal{B}^{\text{on}}|$.

Load Weighted Framework does not require extra information as load variations $L_{b,t}$ are already obtained for the determination of the number of the active RRHs. $L_{b,t}$ yields w_i and the locations of active RRHs yield x_i, both of which form the rest of traffic information exploited. *Load Weighted Framework* can compensate the loss of UE location information in some scenarios. The pseudo-code for *Load Weighted Framework* is given in **Algorithm 2** with changes made based on *Complete Framework*.

Algorithm 2 Load Weighted Framework (Changes)

3: Do nothing

18: $\mathcal{X} \leftarrow \{h_b \in \mathcal{H} : B_b \in \mathcal{B}^{\text{on}}\}$

23: $C^{(l)} \leftarrow \{h_b \in \mathcal{H} : B_b \in (\mathcal{B}^{\text{on}} \cup \mathcal{B}^{\text{off}}_{k-|\mathcal{B}^{\text{on}}|})\}$

Go to Operation 41

30: $\mathcal{W} \leftarrow \{\int_0^T \frac{L_{b,t}}{T} \mathrm{d}t : B_b \in \mathcal{B}^{\text{on}}\}$

35: $v_j^{(l)} \leftarrow (\sum_{x_i \in \mathcal{S}_j^{(l)}} x_i w_i)/(\sum_{w_i \in \{w_i : x_i \in \mathcal{S}_j^{(l)}\}} w_i)$

7.3.3 Energy Weighted Framework

Due to the loss of UE location information and the utilisation of a rough method of traffic quantification, the locations of the active RRHs generated by *Load Weighted Framework* are sometimes inaccurate. To compensate for the loss of the information, power of the uplink transmission bands of the C-RANs from the ambient background of each RRH can be measured and recorded in a clustering period. Each RRH can calculate the integral of the discrete power samples with respect to time at the end of the clustering period, obtaining

$$E_b = \int_0^T P_{b,t}^{\mathrm{r}} \mathrm{d}t, \tag{7.10}$$

where $P_{b,t}^{\mathrm{r}}$ is the received uplink power of B_b sampled at time t. Considering that UE transmission power attenuates rapidly versus distance, E_b can represent the uplink energy virtually accumulated from the links in its local area during a period of time. Therefore, a large number of local links results in more energy accumulated, which indicates a high traffic level in its local area.

This strategy needs all RRHs to sample the uplink power at all times and report the quantified local traffic in the format of energy at the end of a clustering period, which require sleep modes to be designed with more functionality and may lead to a little higher power consumption. Alternatively, the accumulated energy can be realised with extra practical hardware modules such as energy harvesters. RF energy harvesters usually consist of passive receivers and power management modules to convert ambient RF power to stored electricity or power low-power devices. The application of RF energy harvesting has been widely considered for low power devices, such as wireless sensors [100], and may be generalised to cellular network devices [101]. With an energy harvester attached at each RRH, E_b is effectively the uplink energy scavenged at B_b during a clustering period.

Taking the same weighting strategy as *Load Weighted Framework*, $\mathcal{E} = \{E_b\}$ is used for weighting the locations of the RRHs, $\mathcal{H}$, which are treated as $\mathcal{X}$. Therefore, it is named *Energy Weighted Framework*, where the weights of the data entries exactly take the value of the energy virtually accumulated or harvested at the corresponding RRHs, at the beginning of a clustering process. Similarly, the computation of virtual cluster centroids in Equation (7.9) is biased by the assigned weights. With more data entries enriching the information for the update of cluster centroids in every iteration, the final locations of active RRHs are usually more accurate than that in *Load Weighted Framework*.

Although this solution needs additional functionality of sleep modes or extra hardware modules, the most parts of the RRH hardware can still be deactivated and the energy calculated or harvested only needs to be reported when required at the end of a clustering period, which only introduce minor system overheads and power consumption. The traffic information exploited is similar to *Load Weighted Framework*, but with x_i from the locations of all RRHs and w_i yielded by $\mathcal{E}$. The changes of the pseudo-code based on that of *Complete Framework* is given in **Algorithm 3**.

Algorithm 3 Energy Weighted Framework (Changes)

3: Report the energy virtually accumulated or harvested to the central controller when required

9: Report the energy virtually accumulated or harvested to the central controller when required

18: $\mathcal{X} \leftarrow \{h_b \in \mathcal{H}\}$

30: Require $\mathcal{B}$ to report $\mathcal{E} = \{E_b\}$, $\mathcal{W} \leftarrow \mathcal{E}$

35: $v_j^{(l)} \leftarrow (\sum_{x_i \in \mathcal{S}_j^{(l)}} x_i w_i)/(\sum_{w_i \in \{w_i : x_i \in \mathcal{S}_j^{(l)}\}} w_i)$

7.3.4 Random Framework

Considering the simplest strategy without local traffic quantification, $\mathcal{X}$ is then not needed to calculate new cluster centroids. Therefore, cluster centroids are not determined via sequential iterations. After the initial cluster centroids are determined by the same method in *Complete Framework* if the number of active RRHs are to be changed, the clustering process ceases at this point and outputs the initial cluster centroids as the final cluster centroids, i.e. $\mathcal{C}^{(0)}$ are regarded as the locations of the active RRHs for the state adjustment. The rest of the scheme remains unchanged and it is named *Random Framework* due to the random choices of cluster centroids. The pseudo-code is similar to that of *Complete Framework* except the change listed in **Algorithm 4**.

Algorithm 4 Random Framework (Changes)

3: Do nothing

18: Do nothing

30: **Go to** Operation 41

44: Do nothing

In this framework, the central controller only needs to know the locations of the RRHs deployed in the C-RAN, the same as all the other proposed frameworks. The traffic information exploited is $L_{b,t}$ only. It requires the least computational resource and information, but can only guarantee just enough active RRHs. Moreover, it relies on randomness to place active RRHs, which loses the advantage of placing active RRHs in areas with high service demands. It is presented as a benchmark because it still features the power consumption reduction by sleep mode operation at the RRH side. In terms of reducing UE transmission power, *Load Weighted Framework* can be considered as long as the RRH load information is available at the central controller, which is usually the case.

7.4 Performance Evaluation

The system level performance of the proposed Hotspot-oriented Green Frameworks are evaluated using the models introduced in Chapter 3. The SCN is modelled on a square area of 100 metres side length with non-ideal spatial distributions, where λ_{p} of the MHPP is set to $20000/\mathrm{km}^2$ and δ is configured to 5 metres. The RRH power consumption model is adopted for the evaluation of network power consumption.

The arguments of the frameworks are varied and appropriate values are chosen by observing the simulation results, where the QoS should be maintained similar to the system without sleep mode operation. One of the frameworks is compared with a recent baseline strategy [102] that deals with hotspot areas. The performance in reduction in area power consumption is then evaluated and analysed. The average UE transmission power and the generated system overheads of the frameworks are also investigated.

7.4.1 Comparisons with the Baseline Strategy

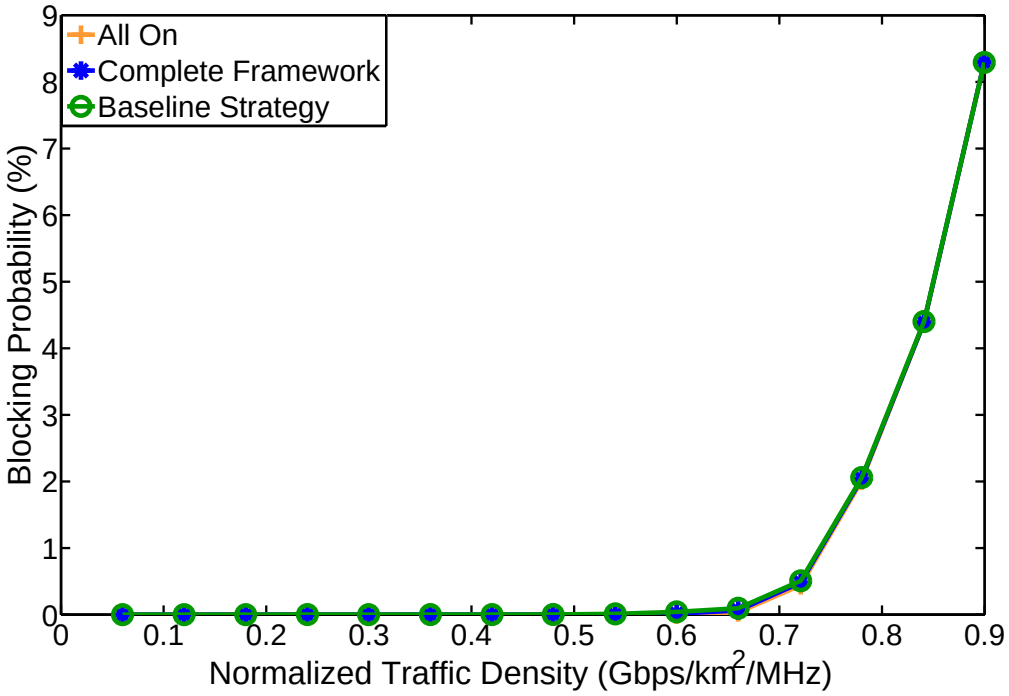
Complete Framework is compared with a recent baseline scheme, a strategic sleeping scheme for the scenario with no UE movements presented in [102], which relies on idealised models to optimise energy efficiency for determining the number of active nodes N_{on}. To deal with hotspot areas, it requires all RRHs to be temporarily active when selecting active RRHs, which is based on the number of covered UEs with qualified SINR. As an analytical approach, it does not consider practical implementation issues, e.g. the frequency of optimising node states. When transplanting it to the considered ultra-small cell C-RAN, it is assumed to run the optimisation every T seconds, the same as the proposed frameworks. Due to the complex

models used in the simulation and the radio resource management in a different scenario, the optimisation of N_{on} in [102] is not applicable and is chosen to be equal to k. By doing this, the comparison can benefit from the controlled system variables and be focused on the mechanism dealing with hotspot areas.

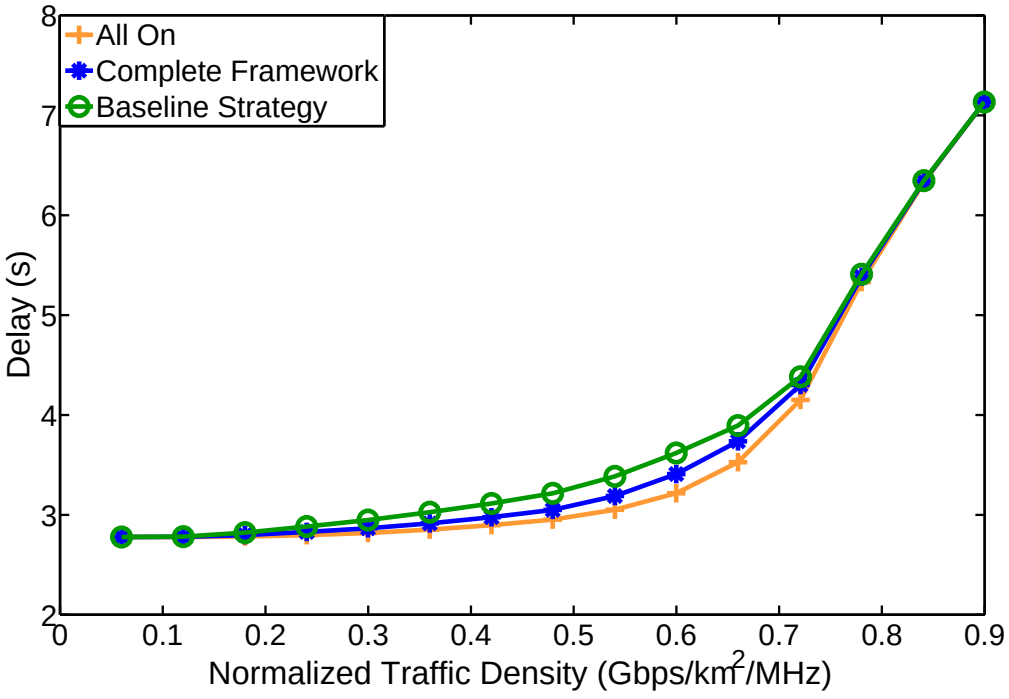
To first ensure QoS, *Complete Framework* and the baseline strategy are varied by setting $L_{\text{ref}} = 6\%$ and $T = 20$ s to have the same blocking probability and similar delay compared to the system without sleep mode operation as shown in Fig. 7.2. Although with the same average proportion of active RRHs, $\overline{\eta}$, by setting the same $L_{\text{ref}} = 6\%$, *Complete Framework* outperforms the baseline strategy in delay because it better manages hotspot areas by putting active RRHs closer to them. This advantage provides more radio resource to areas in demand and yields a higher probability that a UE can acquire radio resources with higher SINRs, which accelerates the transmission rates and yields less delay.

The above superiority is echoed in the average UE transmission power as presented in Fig. 7.3 (a), where *Complete Framework* reduces the UE transmission power by 70% (5.3 dB difference) at low traffic levels compared with the baseline strategy. The average path loss between UEs and their serving RRHs is reduced due to the same reason, which creates lower uplink transmission power on condition of open-loop power control and prolongs the battery lifetimes of mobile devices in a C-RAN with sleep mode operation. Although UE transmission power is increased compared with the no sleep mode operation, it is already very low when introducing ultra-small cells to C-RANs and increased power consumption at the UE side is minor compared with the large reduction in power consumption at the RRH side. As can be seen from Fig. 7.3 (b), the average area power consumption stemming from the RRHs can be reduced more than 79% at traffic levels below 0.1 Gbps/km^2/MHz compared with the C-RAN with no sleep mode operation. The dashed line is the lower bound of the reduction in the total RRH power consumption, where all RRHs are turned into sleep modes.

In Fig. 7.3 (b), the baseline strategy consumes slightly higher power even with the same $\overline{\eta}$, this stems from the mechanism that all RRHs are required to be temporarily switched on to count the covered UEs in their local areas when the central controller is to decide the selection of active RRHs. As energy consumption per switch-on is suggested to be quantified, it is considered when calculating area power consumption and its proportion is shown in Fig. 7.4 (a), where the effect becomes severe at lower traffic levels and causes non-negligible wasted power. *Complete Framework*, however, does not need sleeping RRHs to execute any



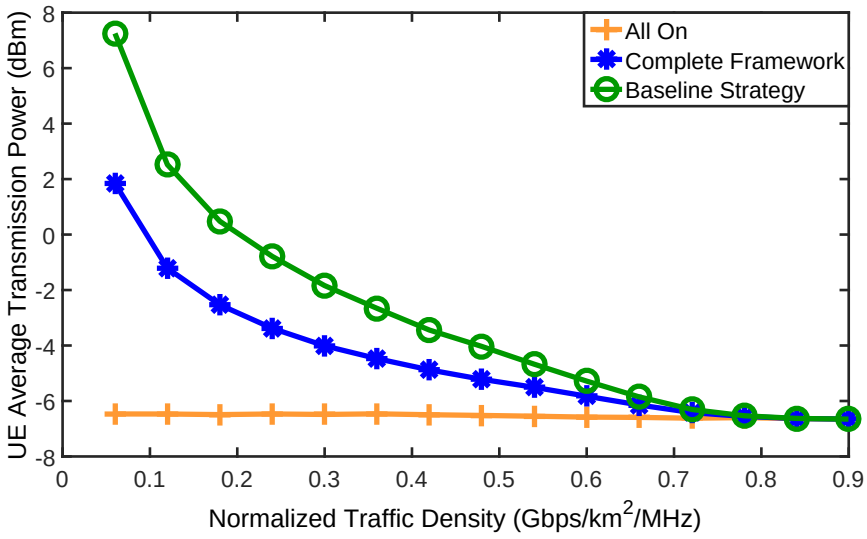
(a)



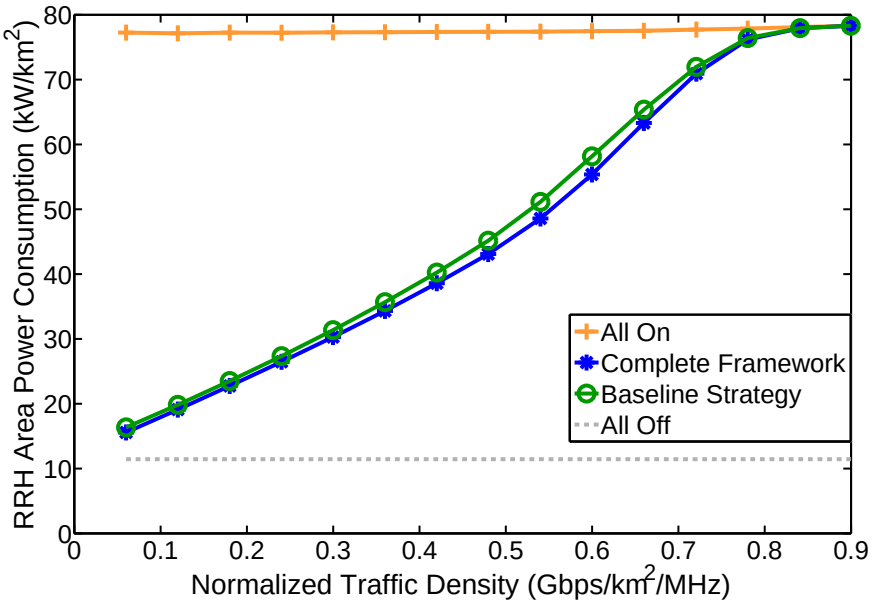
(b)

Figure 7.2: The comparison of QoS between *Complete Framework* and the baseline strategy. (a) Average blocking probability. (b) Average delay.

other operations. It keeps the switching on-off frequency relatively low and stable as revealed in Fig. 7.4 (b). Therefore, it does not pose significant extra area power consumption and system overheads accompanying the switching on-off behaviours.



(a)

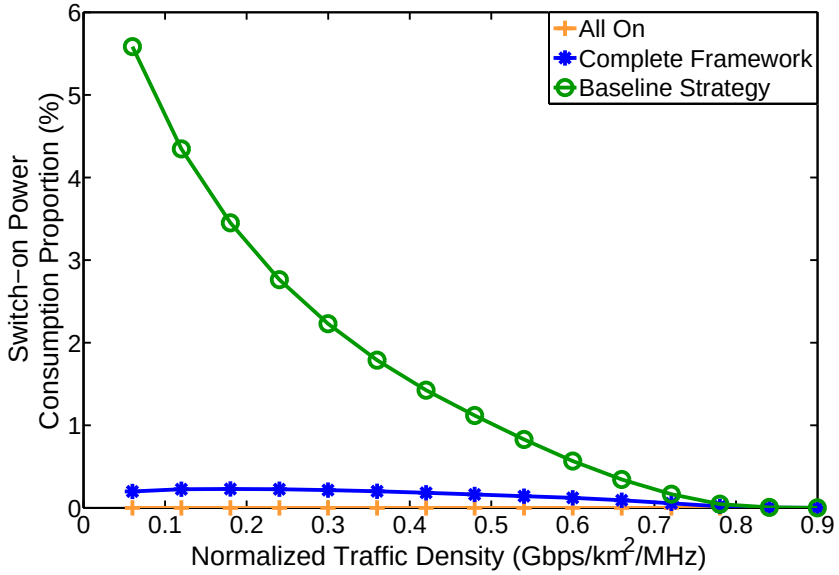


(b)

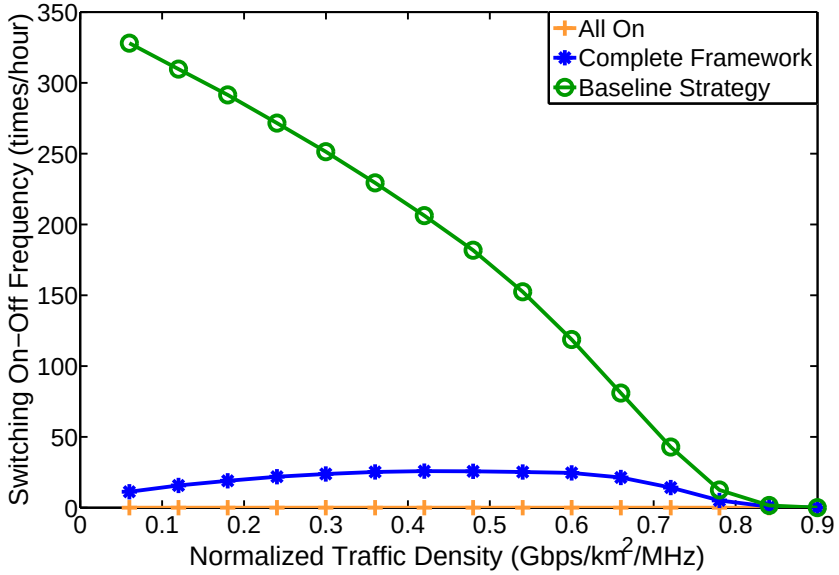
Figure 7.3: The comparison of power consumption between *Complete Framework* and the baseline strategy. (a) Average UE transmission power. (b) Average area power consumption.

7.4.2 Trade-offs in Framework Implementation

In each of the proposed Hotspot-oriented Green Frameworks, only two parameters, L_{ref} and T, are to be determined. The fundamental objective of sleep mode operation is to maintain QoS while switching off some RRHs to save energy. If there are more active RRHs in the C-RAN, there are more radio resources in a certain area, which may lead to better QoS, but less power reduction. The RRH load reference, L_{ref}, is a parameter that can be selected to balance the trade-off. Fig. 7.5 and Fig. 7.6 show the trade-off between delay and area power



(a)



(b)

Figure 7.4: The comparison of switching effects between *Complete Framework* and the baseline strategy. (a) Average proportion of the switch-on power consumption. (b) Average switching on-off frequency.

consumption, selecting *Load Weighted Framework* and $T = 20$ s as examples. As the blocking probability is maintained the same as the system with all RRHs active, it is not shown.

In Fig. 7.5, when L_{ref} is decreased, the delay result can be improved to be the same as the C-RAN without sleep mode operation, which is the upper bound of the QoS performance applying the proposed frameworks. The improvement is because the lower L_{ref} indicates a lower load level per RRH, which brings about more available spectrum resources at each RRH. The probability of acquiring the spectrum resource of a higher SINR for an admitted

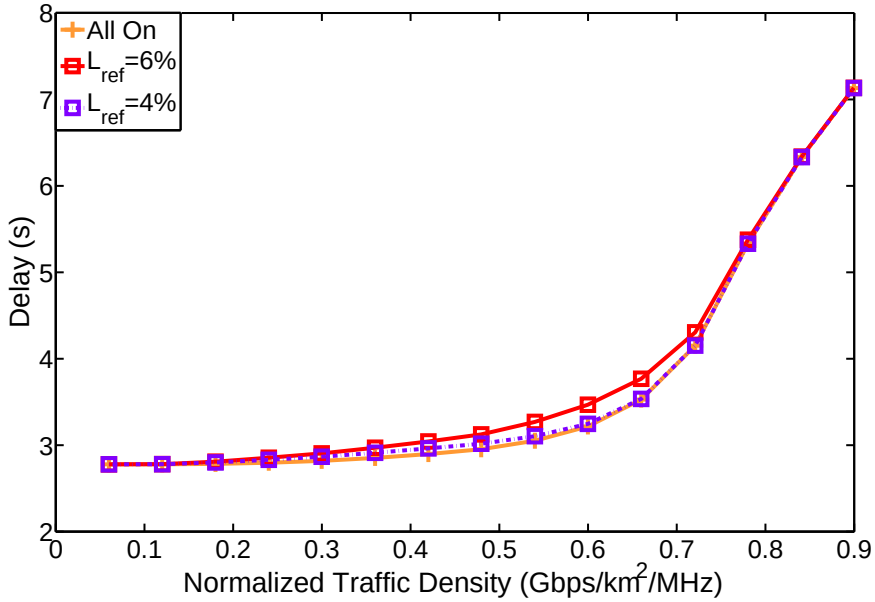


Figure 7.5: The impact of different L_{ref} in average delay for *Load Weighted Framework* at T = 20 s.

UE is usually higher in this case, yielding a higher transmission rate and lower delay.

On the other hand, the average area power consumption contributed by RRHs drawn in Fig. 7.6 shows that a little sacrifice in delay with a slightly higher L_{ref} and fewer active RRHs in the C-RAN benefit from a large reduction in area power consumption especially at medium traffic levels. Therefore, the trade-off between QoS and the power consumption reduction is clear that the QoS improvement brought by keeping more active RRHs in a C-RAN for a certain traffic level is not apparent compared with the power consumption reduction brought by just enough active RRHs.

Another parameter, the clustering period, T, determines the frequency of processing clustering to adjust the RRH states in a C-RAN. If a smaller T is chosen, the switching on-off times of a RRH in a certain period of time is increased as shown in Fig. 7.7 (taking *Load Weighted Framework* with L_{ref} = 6% as an example). This enables a C-RAN to adapt to fast variations of network load levels and traffic distributions, but may yield more overheads. The advantage of fast responses is reflected in reducing the average UE transmission power, which is shown in Fig. 7.8. A more frequent adjustment of RRH states makes the central controller exploit more timely traffic information and increases the probability of putting active RRHs at the current hotspot areas.

To ensure a fair comparison among the proposed four frameworks, they are all simulated by setting L_{ref} = 6% and T = 20 s. The main compromise is that better traffic quantification

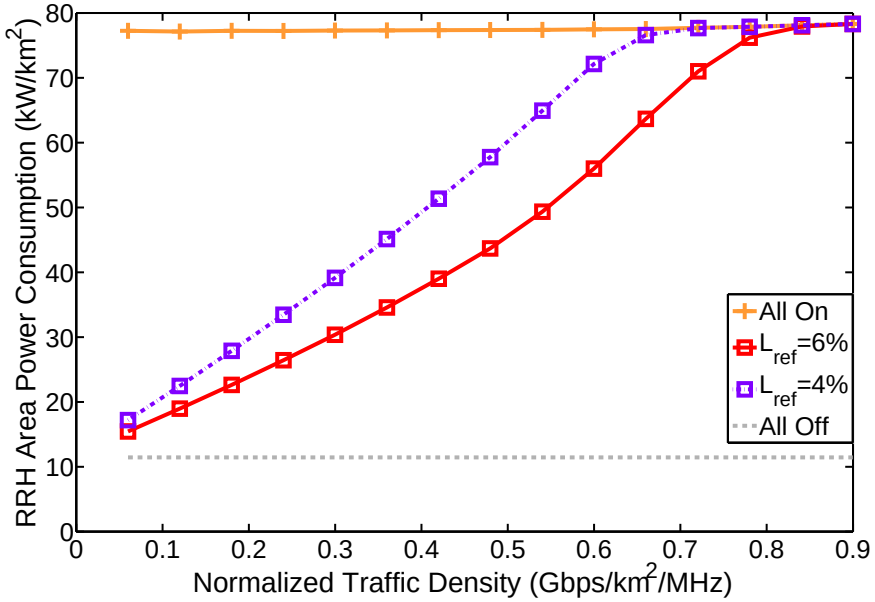


Figure 7.6: The impact of different L_{ref} in the average area power consumption contributed by RRHs for *Load Weighted Framework* at T = 20 s.

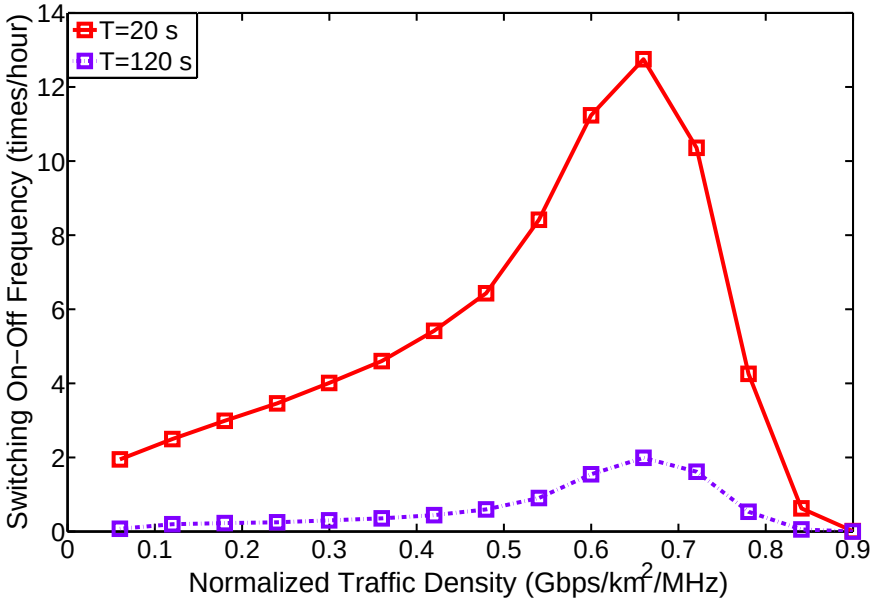


Figure 7.7: The impact of different T in average RRH switching frequency for *Load Weighted Framework*.

based on richer information availability yields more appropriate locations of active RRHs, reducing UE transmission power and better meeting local service demands, but has a higher requirement in terms of either information or infrastructure availability.

For *Complete Framework*, a significant amount of UE location information is needed, which provides the most accurate deduction of hotspot areas. This can be seen from Fig. 7.9, where *Complete Framework* contributes the largest reduction in UE transmission power.

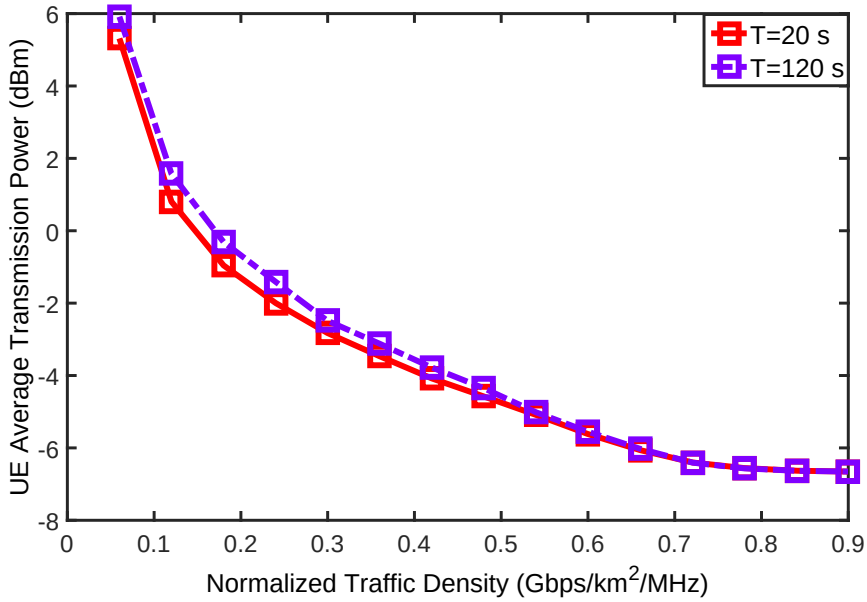


Figure 7.8: The impact of different T in average UE transmission power for *Load Weighted Framework*.

If such UE locations are not supplied or not reliable, the weakness can be compensated to some extent by enabling sleep modes to have the function of measuring uplink power or deploying energy harvesters at RRHs as in *Energy Weighted Framework*, which achieves the second best UE transmission power reduction. Another promising solution requiring no extra information is *Load Weighted Framework*, which performs better than *Random Framework* thanks to the weighting strategy using RRH load levels. The dashed line is plotted based on Equation (7.3) presented before and the average proportion of active RRHs at the corresponding traffic levels, which predicts the average UE transmission power by assuming an uniform active RRH selection. It should be close to the results of *Random Framework*, which adopts random active RRH selection. The difference from the analytical prediction comes from the minor approximation deviation of the distance distribution function used and the association policy assumed that is based on the smallest path loss instead of the Euclidean distance. The impact of different hotspot handling mechanisms in QoS is parallel to UE transmission power that a better active RRH placement renders better delay. This is similar to what is demonstrated before and thus neglected.

7.5 Conclusion

In this chapter, the problem of hotspot areas is analysed. The Hotspot-oriented Green Frameworks involving sleep mode operation are proposed for ultra-small cell C-RANs as

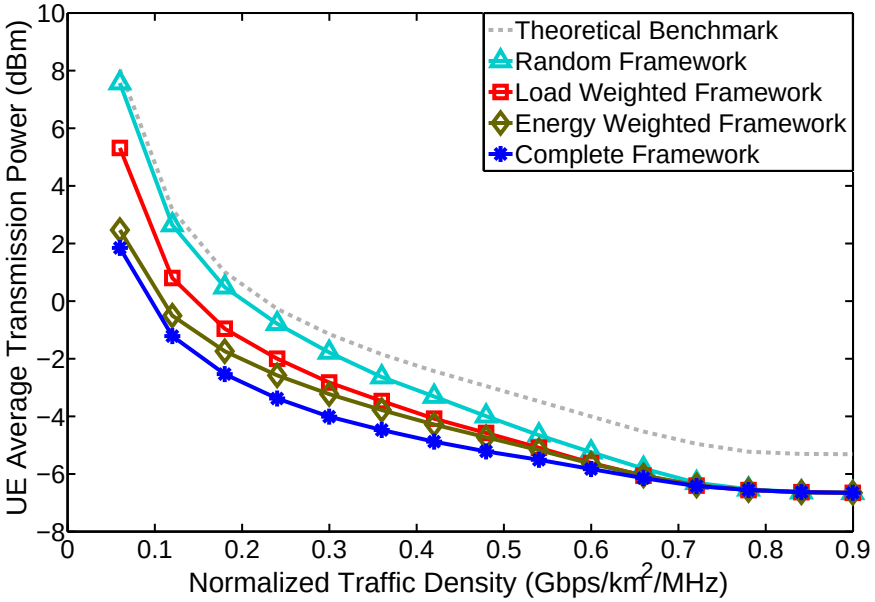


Figure 7.9: The impact of different frameworks in the average UE transmission power.

a centralised solution, filling the gap left by the previous two chapters. The frameworks employ clustering techniques to choose active RRHs based on distinct formats of quantified local traffic for different information availabilities and infrastructures. The Hotspot-oriented Green Frameworks aims to provide global on-demand radio resources to reduce area power consumption while placing active RRHs at hotspot areas to reduce UE transmission power. The simulation results show that one of the frameworks can reduce area power consumption by more than 79% at low traffic levels compared with no sleep mode operation and over 70% in UE transmission power compared with a baseline strategy when the traffic level is below 0.1 Gbps/km^2/MHz. Implementation adaptability of the frameworks are analysed with the trade-offs presented.

Chapter 8

Conclusions, Contributions and Further Work

Contents

8.1 Conclusions

This thesis has investigated the application of sleep mode operation for on-demand radio resource provisioning to alleviate the pressure of the increasing network power consumption brought by the deployment of ultra-small cell networks. Three sleep mode operation schemes have been proposed for various potential circumstances, where SCNs have different architectures and information availabilities. Among them, Adaptive Traffic Perception (ATP)

based sleep mode operation has innovated adaptive monitoring areas to quantify the traffic information in distributed SCNs, which significantly reduces the average network power consumption while maintaining the QoS for different BS layouts. The Traffic-aware Cell Management (TACM) scheme has also been designed for distributed SCNs using a novel approach, adopting Direction of Arrival (DOA) information. It is able to mitigate the side effects introduced by the general sleep mode operation. Hotspot-oriented Green Frameworks have been proposed for centralised SCNs, which optimise the placement of active Remote Radio Heads (RRHs) by a clustering technique, thereby minimising the user transmission power consumption. The analysis of the schemes has been supported by the simulation and the mathematical approximations of the distance distributions proposed.

The research has been guided by the hypothesis mentioned in Chapter 1 as

> *Appropriate traffic information can be exploited in sleep mode operation to improve on-demand radio resource provisioning for ultra-small cell networks of different architectures.*

The hypothesis has indicated the requirements of designing sleep mode operation schemes in different SCN architectures, which are mainly distinguished in terms of traffic information availabilities and the type of decision making engine. This thesis has considered a number of different SCN scenarios, where corresponding sleep mode operation schemes have been proposed, achieving the requirements, respectively. The specific characteristics of SCNs have been appropriately captured with a suitable spatial model to support both simulation-based and theoretical evaluations. The specific demonstration of the hypothesis is given in the following contexts.

One of the modelling techniques of SCNs was first investigated in Chapter 4 to support the evaluation of the sleep mode operation. A stochastic model, type II Matérn Hard-core Point Process (MHPP) has been studied for modelling spatial distributions of BSs of ultra-small cell networks in general cases, forming the basis of the simulation of sleep mode operation. The derivation and approximation of the distance distributions of MHPP have also been proposed, enabling theoretical probabilistic analysis of sleep mode operation proposed in the thesis. Next in Chapter 5, a novel fast-reaction and protocol-friendly sleep mode operation algorithm involving the concept of ATP was proposed for distributed SCNs. It enables each BS in a SCN to create an adaptive monitoring area, where the number of active links is monitored. Together with information on the load and the distance to the nearest

active BS in the power form, the number of active links is used as the traffic information for decision making of a BS in sleep mode operation. The sleep mode operation was further improved in terms of power consumption of sleep modes by the TACM scheme proposed in the following chapter. The TACM scheme adopts DOA information to obtain awareness of the directions to hotspot areas, based on which cell actions including cell division, cell migration and cell death are controlled. DOAs and BS loads play the role as the traffic information in decision making processes of active BSs, which eliminates the necessity of computation at sleeping BSs and improves the on-demand provisioning of radio resources to hotspot areas. In Chapter 7, the proposed green frameworks for centralised architectures were evaluated on Cloud Radio Access Networks (C-RANs). The frameworks involve clustering techniques, which are dedicated to the optimisation of the placement of active RRHs. The choices of active RRHs are a vital part of improving on-demand provisioning to hotspot areas and reducing the user transmission power. In the clustering processes, different formats of traffic information are utilised in different frameworks for C-RANs with specific infrastructure and information availabilities. The load information is also adopted in sleep mode operation. The effectiveness of the proposed schemes has been demonstrated through simulation, with theoretical analysis based on the proposed distance distributions and the comparison with the state-of-the-art, showing significant comprehensive improvements in various aspects including network power consumption, system overhead and user transmission power. Thus, the hypothesis has been demonstrated.

8.2 Novel Contributions

The thesis has proposed, analysed and evaluated sleep mode operation schemes in depth for on-demand radio resource provisioning. The analysis and evaluation of sleep mode operation in the thesis are based on both simulation and the theoretical approaches, the latter of which has not been studied in the state-of-the-art. The details of the original contributions are provided in this section.

8.2.1 Mitigation of the Side Effects of Sleep Mode Operation

Excluding the trade-offs between QoS and reduction in network power consumption, the side effects of sleep mode operation include the additional system overhead and the increased user transmission power. The extra system overhead is generated from the information exchange

and the switching of BSs. The transmission power is raised because of the increasing average distance between transmitters and receivers and is critical at the user side due to their limited battery lives. These issues have not been considered together in the existing sleep mode operation schemes, some of which only have limited consideration in reducing the system overhead generated by switching on-offs. In the thesis, the proposed sleep mode operation scheme in Chapter 6 has been dedicated to this purpose, and is demonstrated to significantly reduce the average user transmission power and the system overhead. The average user transmission power can also be optimised in a centralised SCN as proposed in Chapter 7, which has been demonstrated to be able to reduce the average user transmission power by different extents, depending on the information availability.

8.2.2 Traffic-aware Cell Management

The TACM scheme has been proposed to specifically tackle the side effects associated with sleep mode operation for distributed SCNs in Chapter 6. It has been designed to mitigate the side effects of sleep mode operation by exploiting the application of DOA information, which has never been considered for sleep mode operation in the literature. The scheme has also been dedicated to lowering power consumption of sleep modes by releasing the necessity of computation at a sleeping BS, making it desirable for sleep modes with less functionality. It has been compared with a recent algorithm having the same feature, demonstrating its significant improvement in side effect mitigation, which also mitigates the increase in power consumption generated by the side effects. The scheme is particularly valuable in distributed SCNs with multiple antennas, where DOA information is likely to be available.

8.2.3 Hotspot-oriented Green Frameworks

Hotspot-oriented green frameworks have been proposed in Chapter 7 to address the problem of hotspot areas in centralised SCNs using clustering techniques. Clustering has been adopted to group wireless nodes for various purposes in the literature, but there is no known application in traffic information quantification for sleep mode operation. With a clustering technique and the centralisation advantage, the choice of active RRHs can be optimised from a global view, improving on-demand provisioning of radio resources to hotspot areas and thus reducing the average user transmission power. One of the frameworks has been compared with a recent scheme in the literature, which also attempts to deal with hotspot areas. It

has been demonstrated that the average user transmission power and the system overhead have been significantly reduced and network power consumption has been further decreased. Multiple infrastructure and information availabilities have been considered for centralised SCNs, making the green frameworks applicable in different scenarios.

8.2.4 ATP-based Sleep Mode Operation

The ATP-based sleep mode operation algorithm proposed in Chapter 5 for distributed SCNs has been designed to be compatible with the current LTE protocol. It exploits existing control signals compared with schemes in the literature, which necessitate extra information exchange and introduce extra signalling requirements in order to function. The innovative concept of Adaptive Traffic Perception is that it creates adaptive monitoring areas, which are built on broadcast signals and no point-to-point communication is needed between BSs, realising the necessity of extra signalling and message exchange. It has been demonstrated that the ATP-based sleep mode operation algorithm is able to achieve the fundamental objective of sleep mode operation, maintaining the QoS while reducing the average network power consumption. It has also been evaluated on different BS layouts, shown to maintain the same QoS as the systems without sleep mode operation and the same reduction in power consumption, which indicates its applicability in SCNs having various spatial distributions of BSs.

8.2.5 MHPP Distance Distributions

For a SCN modelled by a stochastic spatial model, the probability of having a certain distance from a user to a BS or from a BS to a BS is captured by distance distribution functions. The distance distributions are essential to obtain the average global system performance in a theoretical analysis. For MHPP adopted for spatial modelling of BSs in the thesis, its nth nearest distance distributions have not been studied well in the literature. The existing closed-form approximations for $n = 1$ are of low accuracy, while there is no approximation for $n \geq 2$. The high-accuracy nearest distance distributions in terms of the contact distribution and the nearest-neighbour distribution have been derived and approximated for the first time in Chapter 4. The resulting cdf functions are straightforward and tractable to be directly used for any further analysis involving distance. The approximations of the nth ($n \geq 2$) distance distributions have also been proposed based on the insights of MHPP. The proposed approximations have been shown to have high accuracy, which guarantees a low

error contributed by distance distributions in any further analysis.

8.2.6 Probabilistic Analysis of Sleep Mode Operation

With stochastic models applied to capture the BS distribution in SCNs, it is feasible to analyse the system performance or the scheme-specific metrics related to sleep mode operation using a probabilistic approach. The theoretical analysis can reveal the quantitative relationship between system metrics and help understand the mechanism of a specific scheme. It also yields hard bounds of system performance and provides a benchmark for system performance. However, this technique has only been used for analysing general system metrics (such as outage probability, interference, etc) in the literature, without the extension to the analysis of sleep mode operation. In the thesis, probabilistic analysis has been applied to the ATP-based sleep mode operation in Chapter 5 with an elegant single equation obtained based on the approximation of a distance distribution function proposed in Chapter 4. It has quantitatively described the correlation between average delay, network power consumption, offered traffic and a scheme-specific controlling parameter. It reveals how the scheme exploits the traffic information to control the sleep mode operation for on-demand radio resource provisioning. The average user transmission power, as an important metric related to sleep mode operation, has been derived in Chapter 7 using probabilistic analysis and appropriate models. A benchmark has also been indicated, giving a hard bound of the average user transmission power to any sleep mode operation scheme.

8.3 Further Work

This section presents recommendations for further work, mainly built on the scope of the thesis. The recommendations include the potential applications and studies of the results provided. They also involve the extension and the transformation of the ideas proposed in the thesis for diverse scenarios.

8.3.1 Applications of Distance Distributions

The derivation and approximation of distance distributions proposed in Chapter 4 are important results for the analysis of wireless networks modelled by MHPPs. Conventionally, the distance distributions are utilised to obtain the mathematical estimations of the metrics

related to distance or path loss. Beyond the sleep mode operation related metrics analysed in the thesis, this analysis could be used in the study of, but are not limited to, outage probability, interference, routing, localisation, cell planning, connectivity, etc. These conventional metrics for HPPP have been studied well, but seldom investigated for MHPP due to the lack of the essential probability generating functional and the distance distributions. Therefore, it is suggested that the approximation of the probability generating functional should be first obtained. Then, the distance distributions of MHPPs should be exploited to acquire theoretical results of the aforementioned metrics with the proposed closed forms of high accuracy.

8.3.2 Hierarchical Clustering

Considering the heterogeneity in SCNs, a group of SCNs of various BS types should be studied if a larger area scenario is investigated. This refers to multi-tier SCNs, where macro BSs provide coverage and small cells of different sizes are responsible for capacity enhancement. Sleep mode operation in multi-tier SCNs is more complicated because of the existence of distinguishing BS types and cell sizes and the coordination between tiers in sleep mode operation remains unknown. Hierarchical clustering is strongly applicable for sleep mode operation of all the BSs in such an architecture, where partially distributed schemes are feasible to be implemented. In hierarchical clustering, the proposed frameworks in Chapter 7 should be transplanted within a single tier. Larger cell BSs of larger coverage areas yield subordination of smaller cell BSs and create a hierarchy on the tiers. Smaller cell BSs will be cluster members of larger cell BSs and the locations of active larger cell BSs can then be determined if the local traffic levels of smaller cell BSs are reasonably quantified and exploited. The strategy of inter-tier management in the hierarchy and the feasibility of information collection should also be investigated.

8.3.3 Distributed Clustering

Although distributed clustering is extensively studied in the field of wireless sensor networks or ad-hoc networks, the known applications to cellular networks and sleep mode operation are limited. Furthermore, the clustered entities considered in the literature are mainly network elements, especially the nodes themselves. However, sleep mode operation is driven by the traffic and it is thus the traffic that should be clustered. This difference makes most dis-

tributed clustering algorithms non-applicable for sleep mode operation. Further applications of clusters in sleep mode operation remain unknown. With the benefits of clustering demonstrated, distributed clustering techniques should be researched for an architecture where centralised management is not possible. The criterion for the judgement of cluster formation should be determined, which may include the distance and the traffic information. The type of traffic information exchanged among neighbour BSs is necessary to be investigated to efficiently form clusters. The mechanism of sleep mode operation managed within a cluster should be studied. A cluster head can be elected to manage the sleep mode operation within a cluster, otherwise BSs can determine their respective states depending on the traffic information collected within the cluster. The optional strategies should be compared in terms of the conventional sleep mode operation metrics and the information exchange overhead.

8.3.4 Handover with Implementation of Sleep Modes

In the research presented in the thesis, handover was assumed to have a simple strategy, where a BS waits until data transmission is finished if it is to switch to a sleep mode. Users and their radio environment were assumed to be static and no handover is processed during data transmission. With sleep mode operation applied, the radio environment within a SCN is less stable. The feasibility of handover when there is a state adjustment of neighbour BSs changing the radio environment should be investigated. It should also be studied whether it is likely to handover users to neighbour active BSs for immediate switching off if a BS decides to sleep. The mobility of users can also be considered, which introduces great demand for appropriate handover especially in SCNs, where the cell sizes are small. This suggested work should be started by investigating the existing handover strategies and understanding how they are affected in SCNs with the implementation of sleep modes. Specific handover strategies should also be designed specifically for SCNs with dynamic topologies. The study should consider SCNs of different architectures and layouts to suit the future potential.

Appendix A

Reduced Palm Distribution

To discuss the Reduced Palm Distribution [29] of a point process Φ, it is necessary to be clear about the probabilities involving conditioning of Φ on the event that Φ contains a point at x. If Y is a configuration set, the event that Φ has property Y is written as $\Phi \in Y$. The aforementioned conditional probability P_x is [29]

$$P_x(Y) = \mathbf{P}(\Phi \in Y \| x) = \mathbf{P}(\Phi \in Y | x \in \Phi). \tag{A.1}$$

The Reduced Palm Distribution $P_x^!$ is therefore

$$P_x^!(Y) = \mathbf{P}(\Phi \setminus \{x\} \in Y \| x). \tag{A.2}$$

For a HPPP, according to the Slivnyak-Mecke Theorem,

$$P_x^! = P. \tag{A.3}$$

The elegant form indicates that the reduced Palm conditioning does not change the distribution of the HPPP.

Appendix B

Retaining Probability of MHPP

Denoting the mark of a primary point x in the parent HPPP Φ_p with intensity λ_p as $m(x)$, the associated mark of an arbitrary $x_0 \in \Phi_\mathrm{p} = \{x_i\} \cup x_0$ $(i \in \mathbb{N}^*)$ is $m(x_0) = t$. Assuming that the random mark t is known, it can be derived that

$$\begin{aligned}
\mathbf{P}(x_0 \in \Phi_\mathrm{m} | x_0 \in \Phi_\mathrm{p}) &= \mathbf{P}\Big(\Phi_\mathrm{m}(B(x_0,\delta) \setminus x_0) = 0 \| x_0\Big) \\
&= \sum_{k=0}^{\infty} \mathbf{P}\Big(\Phi_\mathrm{p}(B(x_0,\delta) \setminus x_0) = k \| x_0\Big) \mathbf{P}\Big(\forall x_i \in \Phi_\mathrm{p} \cap B(x_0,\delta),\ m(x_i) > t\Big) \\
&= \sum_{k=0}^{\infty} \mathbf{P}\Big(\Phi_\mathrm{p}(B(x_0,\delta)) = k\Big) \mathbf{P}\Big(\forall x_i \in \Phi_\mathrm{p} \cap B(x_0,\delta),\ m(x_i) > t\Big) \\
&= \sum_{k=0}^{\infty} \frac{(\lambda_\mathrm{p}\pi\delta^2)^k}{k!} \exp(-\lambda_\mathrm{p}\pi\delta^2)(1-t)^k \\
&= \exp(-\lambda_\mathrm{p}\pi\delta^2) \sum_{k=0}^{\infty} \frac{\left(\lambda_\mathrm{p}\pi\delta^2(1-t)\right)^k}{k!} \\
&= \exp(-\lambda_\mathrm{p}\pi\delta^2) \exp\left(\lambda_\mathrm{p}\pi\delta^2(1-t)\right) \\
&= \exp(-t\lambda_\mathrm{p}\pi\delta^2).
\end{aligned} \tag{B.1}$$

In the derivation, Equation (A.3), Equation (4.1) and the characteristic of an exponential function are utilised. The retaining probability p of the MHPP Φ_m with the hard-core distance δ in a 2D case is equivalent to obtaining the retaining probability of x_0 as

$$p = \int_0^1 \exp(-t\lambda_\mathrm{p}\pi\delta^2)\mathrm{d}t = \frac{1-\exp(-\lambda_\mathrm{p}\pi\delta^2)}{\lambda_\mathrm{p}\pi\delta^2}. \tag{B.2}$$

Appendix C

$k(s)$ for $s > \delta$

The HPPP $\Phi_{\mathrm{p}} = \{z_i\} \cup \{x_0\} \cup \{y_0\}$ $(i \in \mathbb{N}^*)$ with intensity λ_{p} is denoted as the parent point process of the MHPP Φ_{m} with the hard-core distance δ, where x_0 and y_0 are two arbitrary primary points with a pairwise distance s. The associated marks of x_0 and y_0 are assumed to be $m(x_0) = t_x$ and $m(x_0) = t_y$, respectively. B_X is to denote $B(x_0, \delta) \setminus \{x_0\}$ while B_Y is to denote $B(y_0, \delta) \setminus \{y_0\}$.

To derive $k(s)$ for $s > \delta$, $t_x \geq t_y$ can be first assumed. The probability of retaining x_0 knowing $m(x_0) = t_x$ can be obtained based on Equation (B.1) as

$$\mathbf{P}(x_0 \in \Phi_{\mathrm{m}} | x_0 \in \Phi_{\mathrm{p}}) = \exp(-t_x \lambda_{\mathrm{p}} \pi \delta^2). \tag{C.1}$$

Then according to Equation (B.1), the probability of retaining y_0 knowing $m(y_0) = t_y \leq t_x$ conditioning on $x_0 \in \Phi_{\mathrm{m}}$ is

$$\begin{aligned} \mathbf{P}(y_0 \in \Phi_{\mathrm{m}} | x_0 \in \Phi_{\mathrm{m}},\ y_0 \in \Phi_{\mathrm{p}}) &= \mathbf{P}(y_0 \in \Phi_{\mathrm{m}} | \forall z_i \in B_X \cap B_Y,\ m(z_i) \geq t_x \geq t_y) \\ &= \mathbf{P}(\forall z_i \in B_Y \setminus (B_X \cap B_Y),\ m(z_i) \geq t_y) \\ &= \exp\Big(- t_y \lambda_{\mathrm{p}} (V_\delta(s) - \pi \delta^2) \Big), \end{aligned} \tag{C.2}$$

where $V_\delta(s) = v_2(B(x_0, \delta) \cup B(y_0, \delta))$ and $v_2(\cdot)$ denotes the area in a 2D case. Therefore, $k(s)$ in this case is

$$\begin{aligned} &\mathbf{P}(x_0 \in \Phi_{\mathrm{m}},\ y_0 \in \Phi_{\mathrm{m}} | x_0 \in \Phi_{\mathrm{p}},\ y_0 \in \Phi_{\mathrm{p}}) \\ =&\mathbf{P}(x_0 \in \Phi_{\mathrm{m}} | x_0 \in \Phi_{\mathrm{p}}) \mathbf{P}(y_0 \in \Phi_{\mathrm{m}} | x_0 \in \Phi_{\mathrm{m}},\ y_0 \in \Phi_{\mathrm{p}}) \\ =&\exp(-t_x \lambda_{\mathrm{p}} \pi \delta^2) \exp\Big(- t_y \lambda_{\mathrm{p}} (V_\delta(s) - \pi \delta^2) \Big). \end{aligned} \tag{C.3}$$

When considering the case where $t_x < t_y$, $k(s)$ can be obtained by first deriving $\mathbf{P}(y_0 \in \Phi_\mathrm{m} | y_0 \in \Phi_\mathrm{p})$ in a similar way. Then, $\mathbf{P}(x_0 \in \Phi_\mathrm{m} | y_0 \in \Phi_\mathrm{m},\ x_0 \in \Phi_\mathrm{p})$ can be acquired conditioning on the retaining of $y_0 \in \Phi_\mathrm{m}$. $k(s)$ in this case is therefore

$$\mathbf{P}(y_0 \in \Phi_\mathrm{m},\ x_0 \in \Phi_\mathrm{m} | x_0 \in \Phi_\mathrm{p},\ y_0 \in \Phi_\mathrm{p}) = \exp(-t_y \lambda_\mathrm{p} \pi \delta^2) \exp\Big(- t_x \lambda_\mathrm{p} (V_\delta(s) - \pi\delta^2)\Big). \quad \text{(C.4)}$$

The above results are based on the assumption that $m(x_0) = t_x$ and $m(x_0) = t_y$. By deconditioning, $k(s)$ can be derived considering both cases as

$$\begin{aligned} k(s) &= \int_0^1 \int_0^{t_x} \exp(-t_x \lambda_\mathrm{p} \pi \delta^2) \exp\Big(- t_y \lambda_\mathrm{p} (V_\delta(s) - \pi\delta^2)\Big) \mathrm{d}t_y \mathrm{d}t_x \\ &\quad + \int_0^1 \int_0^{t_y} \exp(-t_y \lambda_\mathrm{p} \pi \delta^2) \exp\Big(- t_x \lambda_\mathrm{p} (V_\delta(s) - \pi\delta^2)\Big) \mathrm{d}t_x \mathrm{d}t_y \\ &= \frac{2V_\delta(s)(1 - \mathrm{e}^{-\lambda_\mathrm{p}\pi\delta^2}) - 2\pi\delta^2(1 - \mathrm{e}^{-\lambda_\mathrm{p} V_\delta(s)})}{\lambda_\mathrm{p}^2 V_\delta(s) \pi \delta^2 (V_\delta(s) - \pi\delta^2)}. \end{aligned} \quad \text{(C.5)}$$

Glossary

3GPP	**T**hird **G**eneration **P**artnership **P**roject
4G	**F**ourth **G**eneration
5G	**F**ifth **G**eneration
AI	**A**ntenna **I**nterface
AP	**A**ccess **P**oint
ATP	**A**daptive **T**raffic **P**eception
bps	**b**its **p**er **s**econd
BBU	**B**ase**b**and **U**nit
BPP	**B**inomial **P**oint **P**rocess
BS	**B**ase **S**tation
CAPEX	**C**apital **E**xpenditure
CMOS	**C**omplementary **M**etal–**o**xide–**s**demiconductor
C-RAN	**C**loud **R**adio **A**ccess **N**etwork
CoMP	**Co**ordinated **M**ulti**p**oint
cdf	**c**umulative **d**istribution **f**unction
DOA	**D**irection **o**f **A**rrival
EHF	**E**xtremely **H**igh **F**requency
eNB	**e**volved **N**ode **B**
FAP	**F**emto **A**ccess **P**oint
FBS	**F**emto **B**ase **S**tation
HetSNet	**Het**erogeneous and **S**mall Cell **Net**work
HPPP	**H**omogeneous **P**oisson **P**oint **P**rocess
ICIC	**I**nter-**c**ell **I**nterference **M**anagement
ICT	**I**nformation and **C**ommunication **T**echnology

LOS	**L**ine-**o**f-**s**ight
ISD	**I**nter-**s**ite **D**istance
LTE-A	**L**ong **T**erm **E**volution-**A**dvanced
MAC	**M**edia **A**ccess **C**ontrol
MBS	**M**acro **B**ase **S**tation
MCPP	**M**atérn **C**luster **P**oint **P**rocess
MHPP	**M**atérn **H**ard-core **P**oint **P**rocess
MIMO	**M**ultiple-**I**nput **M**ultiple-**O**utput
MM	**M**obility **M**anagement
MME	**M**obility **M**anagement **E**ntity
mmW	**M**illi**m**etre **W**ave
MSE	**M**ean **S**quared **E**rror
Multi-RAT	**Multi**ple **R**adio **A**ccess **T**echnology
MUSIC	**Mu**ltiple **Si**gnal **C**lassification
NCL	**N**eighbour **C**ell **L**ist
NLOS	**N**on-**l**ine-**o**f-**s**ight
NFV	**N**etwork **F**unction **V**irtualization
PHY	**Phy**sical
QoS	**Q**uality **o**f **S**ervice
OPEX	**Op**erational **Ex**penditure
PA	**P**ower **A**mplifier
PBS	**P**ico **B**ase **S**tation
pdf	**p**robability **d**ensity **f**unction
PPP	**P**oisson **P**oint **P**rocess
RAN	**R**adio **A**ccess **N**etwork
RANaaS	**R**adio **A**ccess **N**etwork **a**s **a** **S**ervice
RAT	**R**adio **A**ccess **T**echnology
RF	**R**adio **F**requency
RN	**R**elay **N**ode
RRH	**R**emote **R**adio **H**ead
RRM	**R**emote **R**esource **M**anagement
SCN	**S**mall **C**ell **N**etwork

SDN	**S**oftware **D**efined **N**etworking
SHF	**S**uper **H**igh **F**requency
SINR	**S**ignal-to-**i**nterference-plus-**n**oise **R**atio
SNR	**S**ignal-to-**n**oise **R**atio
S-GW	**S**erving **G**ate**w**ay
SON	**S**elf-**o**rganised **N**etwork
TACM	**T**raffic-**a**ware **C**ell **M**anagement
UDN	**U**ltra-**d**ense **N**etwork
UE	**U**ser **E**quipment
VRB	**V**irtual **R**esource **B**lock
WLAN	**W**ireless **L**ocal **A**rea **N**etwork
WPAN	**W**ireless **P**ersonal **A**rea **N**etwork

Nomenclature

$\subset$	subset
$\in$	in (set membership)
$\notin$	not in (set membership)
$\cup$	set-theoretic union
$\wedge$	logic and
$\vee$	logic or
$\forall$	for all
$\exists$	there exists
$\mapsto$	map to
$\{\cdots\}$	set-builder
$\lvert\cdots\rvert$	Lebesgue measure
$\binom{\cdot}{\cdot}$	binomial coefficient
$\bar{}$	mean
$\setminus$	relative complement
$\mathbf{1}_A(\cdot)$	indicator function of A
$\mathbf{E}(\cdot)$	expectation
$\mathbf{P}(\cdot)$	probability
$\varnothing$	empty set
$\leftarrow$	get the value of
$\lceil\cdot\rceil$	ceiling
$\lfloor\cdot\rfloor$	floor
$\Vert x$	on condition that x is in the considered set

ρ	inter-site distance of the hexagonal lattice and the square lattice
Φ	spatial point process represented as a countable random set
Φ_{b}	binomial point process
Φ_{m}	Matérn hard-core point process
Φ_{mc}	Matérn cluster point process
Φ_{p}	Poisson point process
γ	hotspot ratio
$(\mu_{\mathrm{X}}, \mu_{\mathrm{Y}})$	coordinate mean of hotspot group centres
σ^2	noise power
σ_{n}	standard deviation of the hotspot user distribution
δ_{h}	minimum distance between hotspot group centres
λ, λ_{f}	file interarrival time
$\overline{\lambda_{\mathrm{f}}}$	mean interarrival time
ε_i, ε_f	delay of file i, delay of file f
$\overline{\varepsilon}$	mean delay
λ_{m}	intensity of Matérn hard-core point process
λ_{p}	intensity of Poisson point process
$\lambda_{\mathrm{u}}(A)$	user equipment density at location A
δ	hard-core distance
β	required signal-to-noise ratio
$\beta_{i,j}^{(t)}$	weight bias of $N_{i,j}$ at time t
$\beta_{i,\max}^{(t)}$	maximum weight bias of $N_{i,j}$ with respect to j at time t
β_{th}	weight bias threshold
$\eta_t(A)$	sleep mode operation at time t and location A denoted as the retaining probability
$\overline{\eta_t}$	assuming uniform $\eta_t(A)$ versus A
$\overline{\eta}$	assuming invariant $\overline{\eta_t}$ versus time
$\Delta\theta$	interval length
α	(Chapter 3) factor considering implementation losses in truncated Shannon bound
α	(Chapter 6) relative angle of the direction of arrival

$\alpha(r)$	distance dependent attenuation
$\psi_i^{(t)}$	memory size of base station i at time t
ψ_{th}	memory size threshold
$\mathcal{A}$	area set
B_m	(Chapter 2) bounded closed set
B_i, B_b, $\mathcal{B}$	base station i, base station b, base station set
B_i^n	nth nearest base station of base station i
$\mathcal{B}^{\text{on}}$	active base station set
$\mathcal{B}_k^{\text{on}}$	subset of $\mathcal{B}^{\text{on}}$ of size k
$\mathcal{B}^{\text{off}}$	sleeping base station set
$\mathcal{B}_{k-\lvert\mathcal{B}^{\text{on}}\rvert}^{\text{off}}$	subset of $\mathcal{B}^{\text{off}}$ of size $k - \lvert\mathcal{B}^{\text{on}}\rvert$
$B(x, R)$	closed disc centred at x with radius R
c	correlation factor
$c_j^{(l)}$	cluster centroid j in iteration l
$\mathcal{C}^{(l)}$	cluster centroid set in iteration l
$d_{i,m}$	distance between base station i and user equipment m
$D_{i,j}$	distance between base station i and base station j
$D_{i,j}$	(Chapter 6) angle of each $N_{i,j}$ relative to B_i
$DivL_{\text{th}}$	cell division load threshold
$DieL_{\text{th}}$	cell death load threshold
E_b	energy (virtually) accumulated by base station b
$\mathcal{E}$	energy set
f_{c}	carrier frequency
$f(r)$, $F(r)$	contact distance distribution (probability density function and cumulative distribution function)
$f_n(r)$, $F_n(r)$	distribution of the distance between an arbitrary point to the nth nearest point in a point process (probability density function and cumulative distribution function)
$F_n^{\text{p},p\lambda_{\text{p}}}(r)$	$F_n(r)$ of the Poisson point process of intensity $p\lambda_{\text{p}}$
$g(r)$, $G(r)$	nearest distance distribution (probability density function and cumulative distribution function)

$g_n(r)$, $G_n(r)$	distribution of the distance between an arbitrary point in a point process to its nth nearest point in the point process (probability density function and cumulative distribution function)
$G_n^{\mathrm{p},p\lambda_\mathrm{p}}(r)$	$G_n(r)$ of the Poisson point process of intensity $p\lambda_\mathrm{p}$
H, H_div, H_die	hysteresis duration, hysteresis duration for cell division, hysteresis duration for cell death
h_b, $\mathcal{H}$	location information of base station b, location information set
I_i	received interference power
$\mathcal{I}_{i,j}$	interval for $W_{i,j}$
$\mathcal{I}_{i,j}^*$	centrally symmetric sector of $\mathcal{I}_{i,j}$
$J^{(l)}$	cost function in iteration l
$k(s)$	probability of retaining two primary points in the thinning process to generate a Matérn hard-core point process
$l(d)$	distance dependent path loss function
L_i, $L_{b,t}$	load of base station i, load of base station b at time t
$\overline{L}$, $\overline{L_b}$	mean network load, mean load of base station b
L_ref	load reference
$\mathcal{L}_\mathrm{h}$	hexagonal lattice
$\mathcal{L}_\mathrm{s}$	square lattice
M_i	equivalent monitoring radius of base station i
M_min	minimum equivalent monitoring radius
M_max	maximum equivalent monitoring radius
$N_{i,j}$, $\mathcal{N}_i$	neighbour base station j of base station i, neighbour base station set of base station i
$N_{i,\mathrm{aim}}$	target $N_{i,j}$ selected by base station i
$\mathcal{N}_i^\mathrm{off}$	sleeping neighbour base station set of base station i
$N(\cdot)$	number of events
$\mathbb{N}$	natural numbers including zero
$\mathbb{N}^*$	natural numbers excluding zero
o	origin

p	retaining probability of the thinning process to generate a Matérn hard-core point process
$P_{i,j}^{(\mathrm{B})}$	power of the received reference signal at base station i from base station j
$P_{i,m,a}^{(\mathrm{U})}$	power of the received reference signal at base station i for link a from user equipment m
P_{ref}	reference signal power
P_{B}	base station active power
P_{S}	base station sleep mode power
$P_{\mathrm{tran}}^{\mathrm{on}}$	transient state power when base stations switch on
$P_{\mathrm{tran}}^{\mathrm{off}}$	transient state power when base stations switch off
$\widehat{P_t^{\mathrm{N}}}$	estimator of mean area power consumption based on simulation results
$\overline{P_t^{\mathrm{N}}}$	mean area power consumption based on probabilistic analysis
$\widehat{P^{\mathrm{u}}}$	estimator of mean user equipment transmission power based on simulation results
$\overline{P_t^{\mathrm{u}}}$	mean user equipment transmission power based on probabilistic analysis
$\widetilde{P_t^{\mathrm{u}}}$	mean user equipment transmission power based on probabilistic analysis assuming a uniform selection of active base stations
r_{e}^F	where intersection of $F_n(r)$ and $F_n^{\mathrm{p},p\lambda_{\mathrm{p}}}(r)$ happens
r_{e}^G	where intersection of $G_n(r)$ and $G_n^{\mathrm{p},p\lambda_{\mathrm{p}}}(r)$ happens
r_{v}^F	radius of the void space for $F_n(r)$
r_{v}^G	radius of the void space for $G_n(r)$
$\mathbb{R}^2$	2-dimensional Euclidean plane
R_i	monitoring reference power of base station i
$R_{\min}$	minimum monitoring reference power
$R_{\max}$	maximum monitoring reference power
S_i	received signal power
$\mathcal{S}_i$	(Chapter 5) memory of base station i
$\mathcal{S}_j^{(l)}$	(Chapter 7) cluster j in iteration l

T	cluster period
$T_{\text{tran}}^{\text{on}}$	transient state time when base stations switch on
$T_{\text{tran}}^{\text{off}}$	transient state time when base stations switch off
$u_{m,a}$	active link a of user equipment m
$v_j^{(l)}$	virtual cluster centroid j in iteration l
w_i	weight of x_i
$W_{i,j}^{(t)}$	weight of neighbour base station j of base station i at time t
W_{th}	size threshold of monitoring lists
$\mathcal{W}$	(Chapter 7) data entry weight set
$\mathcal{W}_i$	(Chapter 5) monitoring list of base station i
$\mathcal{W}_i$	(Chapter 6) weight memory of base station i
x_i	(Chapter 4) generated point of a point process
x_i	(Chapter 7) data entry
$\mathcal{X}$	data entry set

Bibliography

[1] Q. Li, H. Niu, A. Papathanassiou, and G. Wu, "5G network capacity: Key elements and technologies," *IEEE Veh. Technol. Mag.*, vol. 9, no. 1, pp. 71–78, Mar. 2014.

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

[3] R. Wang, H. Hu, and X. Yang, "Potentials and challenges of C-RAN supporting multi-RATs toward 5G mobile networks," *IEEE Access*, vol. 2, pp. 1187–1195, Oct. 2014.

[4] P. Rost, C. Bernardos, A. Domenico, M. Girolamo, M. Lalam, A. Maeder, D. Sabella, and D. Wübben, "Cloud technologies for flexible 5G radio access networks," *IEEE Commun. Mag.*, vol. 52, no. 5, pp. 68–76, May 2014.

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

[6] T. Edler, "Green base stations – how to minimize CO_2 emission in operator networks," in *Bath Base Station Conf. Ericsson Seminar*, 2008.

[7] G. Auer, V. Giannini, C. Desset, I. Godor, P. Skillermark, M. Olsson, M. A. Imran, D. Sabella, M. J. Gonzalez, O. Blume, and A. Fehske, "How much energy is needed to run a wireless network?" *IEEE Wireless Commun.*, vol. 18, no. 5, pp. 40–49, Oct. 2011.

[8] Cisco, "Cisco visual networking index: Global mobile data traffic forecast update, 2013-2018," Cisco, Report, Feb. 2014.

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

[10] M. Shakir, K. Qaraqe, H. Tabassum, M. Alouini, E. Serpedin, and M. Imran, "Green heterogeneous small-cell networks: toward reducing the CO_2 emissions of mobile communications industry using uplink power adaptation," *IEEE Commun. Mag.*, vol. 51, no. 6, pp. 52–61, Jun. 2013.

[11] M. Wildemeersch, T. Quek, C. Slump, and A. Rabbachin, "Cognitive small cell networks: Energy efficiency and trade-offs," *IEEE Trans. Commun.*, vol. 61, no. 9, pp. 4016–4029, Sep. 2013.

[12] Nokia Solutions and Networks. (2015) Technology vision 2020. [Online]. Available: http://resources.alcatel-lucent.com/asset/200270

[13] R. Hu and Y. Qian, "An energy efficient and spectrum efficient wireless heterogeneous network framework for 5G systems," *IEEE Commun. Mag.*, vol. 52, no. 5, pp. 94–101, May 2014.

[14] 3GPP. Heterogeneous networks in LTE. [Online]. Available: http://www.3gpp.org/technologies/keywords-acronyms/98-lte

[15] C.-L. I, J. Huang, R. Duan, C. Cui, J. Jiang, and L. Li, "Recent progress on C-RAN centralization and cloudification," *IEEE Access*, vol. 2, pp. 1030–1039, Sep. 2014.

[16] C. M. R. Institute. (2013, December) C-RAN: The road towards green RAN. [Online]. Available: http://labs.chinamobile.com/cran/wp-content/uploads/2014/06/20140613-C-RAN-WP-3.0.pdf

[17] D. Wübben, P. Rost, J. Bartelt, M. Lalam, V. Savin, M. Gorgoglione, A. Dekorsy, and G. Fettweis, "Benefits and impact of cloud computing on 5G signal processing: Flexible centralization through cloud-RAN," *IEEE Signal Process. Mag.*, vol. 31, no. 6, pp. 35–44, Nov. 2014.

[18] J. Wu, Z. Zhang, Y. Hong, and Y. Wen, "Cloud radio access network (C-RAN): a primer," *IEEE Netw.*, vol. 29, no. 1, pp. 35–41, Jan. 2015.

[19] D. Sabella, P. Rost, Y. Sheng, E. Pateromichelakis, U. Salim, P. Guitton-Ouhamou, M. di Girolamo, and G. Giuliani, "RAN as a service: Challenges of designing a flexible

RAN architecture in a cloud-based heterogeneous mobile network," in *Proc. Future Netw. and Mobile Summit*, 2013, pp. 1–8.

[20] R. Baldemair, T. Irnich, K. Balachandran, E. Dahlman, G. Mildh, Y. Selén, S. Parkvall, M. Meyer, and A. Osseiran, "Ultra-dense networks in millimeter-wave frequencies," *IEEE Commun. Mag.*, vol. 53, no. 1, pp. 202–208, Jan. 2015.

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

[22] S.-P. Yeh, S. Talwar, G. Wu, N. Himayat, and K. Johnsson, "Capacity and coverage enhancement in heterogeneous networks," *IEEE Wireless Commun.*, vol. 18, no. 3, pp. 32–38, Jun. 2011.

[23] S. Navaratnarajah, A. Saeed, M. Dianati, and M. Imran, "Energy efficiency in heterogeneous wireless access networks," *IEEE Wireless Commun.*, vol. 20, no. 5, pp. 37–43, Oct. 2013.

[24] B. Haberland, F. Derakhshan, H. Grob-Lipski, R. Klotsche, W. Rehm, P. Schefczik, and M. Soellner, "Radio base stations in the cloud," *Bell Labs Tech. J.*, vol. 18, no. 1, pp. 129–152, Jun. 2013.

[25] I. Hwang, B. Song, and S. Soliman, "A holistic view on hyper-dense heterogeneous and small cell networks," *IEEE Commun. Mag.*, vol. 51, no. 6, pp. 20–27, Jun. 2013.

[26] A. Anpalagan, M. Bennis, and R. Vannithamby, *Design and Deployment of Small Cell Networks*. Cambridge, UK: Cambridge University Press, Dec. 2015.

[27] H. ElSawy, E. Hossain, and M. Haenggi, "Stochastic geometry for modeling, analysis, and design of multi-tier and cognitive cellular wireless networks: A survey," *IEEE Commun. Surveys Tuts.*, vol. 15, no. 3, pp. 996–1019, Jul. 2013.

[28] A. Guo and M. Haenggi, "Spatial stochastic models and metrics for the structure of base stations in cellular networks," *IEEE Trans. Wireless Commun.*, vol. 12, no. 11, pp. 5800–5812, Nov. 2013.

[29] S. N. Chiu, D. Stoyan, W. S. Kendall, and J. Mecke, *Stochastic Geometry and Its Applications*, 3rd ed., ser. Wiley Series in Probability and Statistics. Chichester, UK: John Wiley & Sons Ltd, Aug. 2013.

[30] M. Haenggi, *Stochastic Geometry for Wireless Networks.* New York, NY, USA: Cambridge University Press, Oct. 2012.

[31] S. Srinivasa and M. Haenggi, "Distance distributions in finite uniformly random networks: Theory and applications," *IEEE Trans. Veh. Technol.*, vol. 59, no. 2, pp. 940–949, Feb. 2010.

[32] L. Wu, Y. Zhong, and W. Zhang, "Spatial statistical modeling for heterogeneous cellular networks - an empirical study," in *Proc. IEEE 79th Veh. Technol. Conf.*, May 2014, pp. 1–6.

[33] F. Baccelli and B. Blaszczyszyn, *Stochastic Geometry and Wireless Networks, Volume II - Applications*, ser. Foundations and Trends® in Networking. Breda, Netherlands: NoW Publishers, 2010, vol. 4, no. 1-2, pp. 1-312.

[34] S.-R. Cho and W. Choi, "Coverage and load balancing in heterogeneous cellular networks with minimum cell separation," *IEEE Trans. Mobile Comput.*, vol. 13, no. 9, pp. 1955–1966, Sep. 2014.

[35] Y. Zhou, Z. Zhao, Q. Ying, R. Li, X. Zhou, and H. Zhang, "Two-tier spatial modeling of base stations in cellular networks," in *Proc. IEEE 25th Int. Symp. Personal Indoor and Mobile Radio Commun.*, Sep. 2014, pp. 1570–1574.

[36] A. Bianzino, C. Chaudet, D. Rossi, and J. Rougier, "A survey of green networking research," *IEEE Commun. Surveys Tuts.*, vol. 14, no. 1, pp. 3–20, Feb. 2012.

[37] K. Davaslioglu and E. Ayanoglu, "Quantifying potential energy efficiency gain in green cellular wireless networks," *IEEE Commun. Surveys Tuts.*, vol. 16, no. 4, pp. 2065–2091, Nov. 2014.

[38] D. Feng, C. Jiang, G. Lim, J. Cimini, L.J., G. Feng, and G. Li, "A survey of energy-efficient wireless communications," *IEEE Commun. Surveys Tuts.*, vol. 15, no. 1, pp. 167–178, Feb. 2013.

[39] D. Lister. (2009, June) An operator's view on green radio. Vodafone Group Research & Development. [Online]. Available: http://www.green-communications.net/icc09/docs/GreenComm-ICC09-Keynote3-Lister.pdf

[40] L. Correia, D. Zeller, O. Blume, D. Ferling, Y. Jading, I. Gódor, G. Auer, and L. Van der Perre, "Challenges and enabling technologies for energy aware mobile radio networks," *IEEE Commun. Mag.*, vol. 48, no. 11, pp. 66–72, Nov. 2010.

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

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

[43] J. Louhi, "Energy efficiency of modern cellular base stations," in *Proc. 29th Int. Telecommun. Energy Conf.*, 2007, pp. 475–476.

[44] H. Claussen, L. Ho, and F. Pivit, "Effects of joint macrocell and residential picocell deployment on the network energy efficiency," in *Proc. IEEE 19th Int. Symp. Personal, Indoor and Mobile Radio Commun.*, 2008, pp. 1–6.

[45] X. Li, X. Zhang, and W. Wang, "An energy-efficient cell planning strategy for heterogeneous network based on realistic traffic data," in *Proc. Int. Conf. Comput. Manage. and Telecomun.*, Apr. 2014, pp. 122–127.

[46] S. Wang, W. Zhao, and C. Wang, "Budgeted cell planning for cellular networks with small cells," *IEEE Trans. Veh. Technol.*, vol. 64, no. 10, pp. 4797–4806, Oct. 2015.

[47] M. Peng, Y. Li, J. Jiang, J. Li, and C. Wang, "Heterogeneous cloud radio access networks: a new perspective for enhancing spectral and energy efficiencies," *IEEE Wireless Commun.*, vol. 21, no. 6, pp. 126–135, Dec. 2014.

[48] L. Budzisz, F. Ganji, G. Rizzo, M. Marsan, M. Meo, Y. Zhang, G. Koutitas, L. Tassiulas, S. Lambert, B. Lannoo, M. Pickavet, A. Conte, I. Haratcherev, and A. Wolisz, "Dynamic resource provisioning for energy efficiency in wireless access networks: A survey and an outlook," *IEEE Commun. Surveys Tuts.*, vol. 16, no. 4, pp. 2259–2285, Nov. 2014.

[49] M. Marsan, G. Bucalo, A. Di Caro, M. Meo, and Y. Zhang, "Towards zero grid electricity networking: Powering BSs with renewable energy sources," in *Proc. IEEE Int. Conf. Commun. Workshops*, 2013, pp. 596–601.

[50] M. A. Marsan, L. Chiaraviglio, D. Ciullo, and M. Meo, "Optimal energy savings in cellular access networks," in *Proc. IEEE Int. Conf. Commun. Workshops*, Jun. 2009, pp. 1–5.

[51] M. A. Marsan, L. Chiaraviglio, D. Ciullo, and M. Meo, "Multiple daily base station switch-offs in cellular networks," in *Proc. 4th Int. Conf. Commun. and Electron.*, Aug. 2012, pp. 245–250.

[52] O. Blume, H. Eckhardt, S. Klein, E. Kuehn, and W. M. Wajda, "Energy savings in mobile networks based on adaptation to traffic statistics," *Bell Labs Tech. J.*, vol. 15, no. 2, pp. 77–94, Sep. 2010.

[53] 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.

[54] I. Ashraf, L. T. W. Ho, and H. Claussen, "Improving energy efficiency of femtocell base stations via user activity detection," in *Proc. IEEE Wireless Commun. and Networking Conf.*, Apr. 2010, pp. 1–5.

[55] M. A. Marsan, L. Chiaraviglio, D. Ciullo, and M. Meo, "Switch-off transients in cellular access networks with sleep modes," in *Proc. IEEE Int. Conf. Commun. Workshops*, Jun. 2011, pp. 1–6.

[56] K. Samdanis, D. Kutscher, and M. Brunner, "Self-organized energy efficient cellular networks," in *Proc. IEEE 21st Int. Symp. Personal, Indoor and Mobile Radio Commun.*, 2010, Conference Proceedings, pp. 1665–1670.

[57] S. Kokkinogenis and G. Koutitas, "Dynamic and static base station management schemes for cellular networks," in *Proc. IEEE Global Commun. Conf.*, Dec. 2012, pp. 3443–3448.

[58] A. Capone, I. Filippini, B. Gloss, and U. Barth, "Rethinking cellular system architecture for breaking current energy efficiency limits," in *Proc. Sustainable Internet and ICT for Sustainability*, Oct. 2012, pp. 1–5.

[59] Z. Li, D. Grace, and P. Mitchell, "Traffic perception based topology management for 5G green ultra-small cell networks," in *Proc. IEEE 1st Int. Workshop Cognitive Cellular Syst.*, Sep. 2014, pp. 1–5.

[60] Z. Li, D. Grace, and P. Mitchell, "Cell division, migration and death for energy efficient 5G ultra-small cell networks," in *Proc. IEEE Globecom Workshops*, Dec. 2014, pp. 942–947.

[61] I. Ashraf, F. Boccardi, and L. Ho, "Sleep mode techniques for small cell deployments," *IEEE Commun. Mag.*, vol. 49, no. 8, pp. 72–79, Aug. 2011.

[62] H. Wang, X. Zhou, and M. Reed, "Coverage and throughput analysis with a non-uniform small cell deployment," *IEEE Trans. Wireless Commun.*, vol. 13, no. 4, pp. 2047–2059, Apr. 2014.

[63] T. Novlan, R. Ganti, A. Ghosh, and J. Andrews, "Analytical evaluation of fractional frequency reuse for OFDMA cellular networks," *IEEE Trans. Wireless Commun.*, vol. 10, no. 12, pp. 4294–4305, Dec. 2011.

[64] M. Haenggi, J. Andrews, F. Baccelli, O. Dousse, and M. Franceschetti, "Stochastic geometry and random graphs for the analysis and design of wireless networks," *IEEE J. Sel. Areas Commun.*, vol. 27, no. 7, pp. 1029–1046, Sep. 2009.

[65] M. Haenggi, "Mean interference in hard-core wireless networks," *IEEE Commun. Lett.*, vol. 15, no. 8, pp. 792–794, Aug. 2011.

[66] V. M. Nguyen and H. Claussen, "Efficient self-optimization of neighbour cell lists in macrocellular networks," in *Proc. IEEE 21st Int. Symp. Personal, Indoor and Mobile Radio Commun.*, 2010, pp. 1923–1928.

[67] D. Kim, B. Shin, D. Hong, and J. Lim, "Self-configuration of neighbor cell list utilizing E-UTRAN NodeB scanning in LTE systems," in *Proc. 7th IEEE Consum. Commun. and Netw. Conf.*, 2010, pp. 1–5.

[68] J. Lim and D. Hong, "Management of neighbor cell lists and physical cell identifiers in self-organizing heterogeneous networks," *J. Commun. and Netw.*, vol. 13, no. 4, pp. 367–376, Aug. 2011.

[69] Z. Becvar, P. Mach, and M. Vondra, "Optimization of SINR-based neighbor cell list for networks with small cells," in *Proc. IEEE 24th Int. Symp. Personal Indoor and Mobile Radio Commun.*, 2013, pp. 2346–2351.

[70] 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.

[71] 3GPP, *Evolved Universal Terrestrial Radio Access (E-UTRA): Physical layer procedures*, 3GPP Standard TS 36.213, Rev. V12.0.0, Dec. 2013.

[72] P. Kyösti, J. Meinilä, L. Hentilä, X. Zhao, T. Jämsä, C. Schneider, M. Narandzić, M. Milojević, A. Hong, J. Ylitalo, V. Holappa, M. Alatossava, R. Bultitude, Y. Jong, and T. Rautiainen, "IST-4-027756 WINNER II D1.1.2 V1.2 WINNER II channel models," Report, Feb. 2008.

[73] 3GPP, *Evolved Universal Terrestrial Radio Access (E-UTRA): Radio Frequency (RF) system scenarios*, 3GPP Standard TR 36.942, Rev. V11.0.0, Sep. 2012.

[74] 3GPP, *Evolved Universal Terrestrial Radio Access (E-UTRA): Further advancements for E-UTRA physical layer aspects*, 3GPP Standard TR 36.814, Rev. V9.0.0, Mar. 2010.

[75] X. Guo, S. Zhou, Z. Niu, and P. Kumar, "Optimal wake-up mechanism for single base station with sleep mode," in *Proc. 25th Int. Teletraffic Congr.*, Sep. 2013, pp. 1–8.

[76] D. Sabella, A. De Domenico, E. Katranaras, M. Imran, M. di Girolamo, U. Salim, M. Lalam, K. Samdanis, and A. Maeder, "Energy efficiency benefits of RAN-as-a-service concept for a cloud-based 5G mobile network infrastructure," *IEEE Access*, vol. 2, pp. 1586–1597, Jan. 2014.

[77] W. Vereecken, I. Haratcherev, M. Deruyck, W. Joseph, M. Pickavet, L. Martens, and P. Demeester, "The effect of variable wake up time on the utilization of sleep modes in femtocell mobile access networks," in *Proc. 9th Annu. Conf. Wireless On-demand Netw. Syst. and Services*, Jan. 2012, pp. 63–66.

[78] B. Matérn, *Spatial Variation*, 2nd ed., ser. Lecture Notes in Statistics. New York, NY, USA: Springer-Verlag, 1986, vol. 36.

[79] B. Cho, K. Koufos, and R. Jantti, "Bounding the mean interference in matérn type ii hard-core wireless networks," *IEEE Wireless Commun. Lett.*, vol. 2, no. 5, pp. 563–566, Oct. 2013.

[80] A. M. Ibrahim, T. ElBatt, and A. El-Keyi, "Coverage probability analysis for wireless networks using repulsive point processes," in *Proc. IEEE 24th Int. Symp. Personal, Indoor and Mobile Radio Commun.*, Sep. 2013, pp. 1002–1007.

[81] M. Haenggi, "On distances in uniformly random networks," *IEEE Trans. Inf. Theory*, vol. 51, no. 10, pp. 3584–3586, Oct. 2005.

[82] J. G. Andrews, F. Baccelli, and R. K. Ganti, "A tractable approach to coverage and rate in cellular networks," *IEEE Trans. Commun.*, vol. 59, no. 11, pp. 3122–3134, Nov. 2011.

[83] E. Oh, K. Son, and B. Krishnamachari, "Dynamic base station switching-on/off strategies for green cellular networks," *IEEE Trans. Wireless Commun.*, vol. 12, no. 5, pp. 2126–2136, May 2013.

[84] R. O. Schmidt, "Multiple emitter location and signal parameter estimation," *IEEE Trans. Antennas Propag.*, vol. 34, no. 3, pp. 276–280, Mar. 1986.

[85] L. C. Godara, "Application of antenna arrays to mobile communications, part II: Beamforming and direction-of-arrival considerations," *Proc. IEEE*, vol. 85, no. 8, pp. 1195–1245, 1997.

[86] A. Hirata, E. Taillefer, H. Yamada, and T. Ohira, "Handheld direction of arrival finder with electronically steerable parasitic array radiator using the reactance-domain MUltiple SIgnal Classification algorithm," *IET Microw., Antennas & Propag.*, vol. 1, no. 4, pp. 815–821, Aug. 2007.

[87] A. Gaber and A. Omar, "Utilization of multiple-antenna multicarrier systems and NLOS mitigation for accurate wireless indoor positioning," *IEEE Trans. Wireless Commun.*, vol. 15, no. 10, pp. 6570–6584, Oct. 2016.

[88] D. Liu, K. Liu, Y. Ma, and J. Yu, "Joint TOA and DOA localization in indoor environment using virtual stations," *IEEE Commun. Lett.*, vol. 18, no. 8, pp. 1423–1426, Aug. 2014.

[89] Z. Wang and S. A. Zekavat, "Omnidirectional mobile NLOS identification and localization via multiple cooperative nodes," *IEEE Trans. Mobile Comput.*, vol. 11, no. 12, pp. 2047–2059, Dec. 2012.

[90] V. Zhang and A. Wong, "Combined AOA and TOA NLOS localization with nonlinear programming in severe multipath environments," in *Proc. IEEE Wireless Commun. and Networking Conf.*, 2009, pp. 1–6.

[91] A. Checko, H. Christiansen, Y. Yan, L. Scolari, G. Kardaras, M. Berger, and L. Dittmann, "Cloud RAN for mobile networks - a technology overview," *IEEE Commun. Surveys Tuts.*, vol. 17, no. 1, pp. 405–426, Mar. 2015.

[92] J. Yu and P. Chong, "A survey of clustering schemes for mobile ad hoc networks," *IEEE Commun. Surveys Tuts.*, vol. 7, no. 1, pp. 32–48, 1st Qtr. 2005.

[93] S. AlMheiri and H. AlQamzi, "MANETs and VANETs clustering algorithms: A survey," in *Proc. IEEE 8th GCC Conf. and Exhibition*, 2015, pp. 1–6.

[94] K. Hosseini, H. Dahrouj, and R. Adve, "Distributed clustering and interference management in two-tier networks," in *Proc. IEEE Global Commun. Conf.*, 2012, pp. 4267–4272.

[95] A. Abdelnasser, E. Hossain, and D. I. Kim, "Clustering and resource allocation for dense femtocells in a two-tier cellular OFDMA network," *IEEE Trans. Wireless Commun.*, vol. 13, no. 3, pp. 1628–1641, Mar. 2014.

[96] R. Estrada, H. Otrok, and Z. Dziong, "Clustering and dynamic resource allocation for macro-femtocell networks," in *Proc. 16th Int. Telecommun. Netw. Strategy and Planning Symp.*, 2014, pp. 1–6.

[97] W. Li, W. Zheng, Y. Xie, and X. Wen, "Clustering based power saving algorithm for self-organized sleep mode in femtocell networks," in *Proc. 15th Int. Symp. Wireless Personal Multimedia Commun.*, 2012, pp. 379–383.

[98] G. Lee, H. Kim, Y.-T. Kim, and B.-H. Kim, "Delaunay triangulation based green base station operation for self organizing network," in *Proc. IEEE Int. Conf. Green Comput. and Commun. and IEEE Internet of Things and IEEE Cyber, Physical and Social Comput.*, Aug. 2013, pp. 1–6.

[99] S. Samarakoon, M. Bennis, W. Saad, and M. Latva-aho, "Dynamic clustering and sleep mode strategies for small cell networks," in *Proc. 11th Int. Symp. Wireless Commun. Syst.*, 2014, pp. 934–938.

[100] H. Visser and R. Vullers, "RF energy harvesting and transport for wireless sensor network applications: Principles and requirements," *Proc. IEEE*, vol. 101, no. 6, pp. 1410–1423, Jun. 2013.

[101] X. Lu, P. Wang, D. Niyato, and E. Hossain, "Dynamic spectrum access in cognitive radio networks with RF energy harvesting," *IEEE Wireless Commun.*, vol. 21, no. 3, pp. 102–110, Jun. 2014.

[102] C. Liu, B. Natarajan, and H. Xia, "Small cell base station sleep strategies for energy efficiency," *IEEE Trans. Veh. Technol.*, vol. 65, no. 3, pp. 1652–1661, Mar. 2016.