Energy-Efficient Algorithms and Access Schemes for Small Cell Networks

Haluk Celebi

Submitted in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy
in the Graduate School of Arts and Sciences

COLUMBIA UNIVERSITY

2020

Abstract

Energy-Efficient Algorithms and Access Schemes for Small Cell Networks

Haluk Celebi

Dense deployment of small base stations (SBSs) brings new challenges such as growing energy consumption, increased carbon footprint, higher inter-cell interference, and complications in handover management. These challenges can be dealt with by taking advantage of sleep/idle mode capabilities of SBSs, and exploiting the delay tolerance of data applications, as well as utilizing information derived from the statistical distributions of SBSs and user equipment (UE)-SBS associations. This dissertation focuses on the formulation of mathematical models and proposes energy efficient algorithms for small cell networks (SCN). It is shown that delay tolerance of some data applications can be taken advantage of to save energy in SCN. This dissertation introduces practical models to study the performance of delayed access to SCNs. Operational states of SBS are modeled as a Markov chain and their probability distributions are analyzed. Also, it argues that SCN can be operated to save energy during low traffic periods by taking advantage of user equipments' (UEs) delay tolerance in SCN while providing high access probability within bounded transmission range.

Dense deployment of SCNs cause an increase in overlapping SBS coverage areas, allowing UEs to establish communication with multiple SBSs. A new load metric as a function of the number of SBSs in UE's communication range is defined, and its statistics are rigorously analyzed. Energy saving algorithms based on aforementioned load metric are developed and their efficiencies are compared. Besides, UE's delay tolerance allows establishing communication with close-by SBSs that are either in fully active mode or in sleeping mode. Improvements in coverage probability and bitrate are analyzed by considering different delay tolerance values for UEs. Key parameters such as UE's communication range are optimized with respect to SBS density and delay tolerance.

The fundamental problem of local versus remote edge/fog computing and its inherent tradeoffs are studied from a queuing perspective taking into account user/SBS density, server capacity and latency constraints. The task offloading problem is cast as an M/M/1(c) queue in which CPU intensive tasks arrive according to Poisson process and receive service subject to a tolerable delay. The higher the proportion of locally computed tasks, the less traffic SCN handles between edge processor and UE. Therefore, low utilization of SCN can be interperted as increased spectral efficiency due to low interference and close UE-SBS distance. Tradeoff between delay dependent SCN utilization and spectral efficiency is evaluated at high and low traffic loads.

Table of Contents

List of Tables

List of Figures

Acknowledgements

Firstly, I would like to express my sincere gratitude to my advisor Prof. Henning Schulzrinne for the continuous support of my Ph.D. study, for his patience, motivation, and immense knowledge. His guidance and key critics helped me complete this thesis.

Besides my adviser, I would like to thank Prof. İsmail Güvenç and Dr. Yavuz Yapıcı for bringing their insightful comments and encouragement, and research ideas which incented me to widen my research from various perspectives.

My sincere thanks also goes to Peter R. Kinget, and Prof. Gil Zussman for their help and support. Without their precious support it would not be possible to complete this dissertation. Last but not least, I would like to thank the staff of Electrical Engineering Department for being always understanding and helpful.

To my parents, Mehmet and Makbule

Chapter 1: Introduction

Energy and capacity have been fruitful research topics that bring both challenges and advantages for wireless networks. Today, due to rising traffic demand, cellular networks are on the crossroads of a major shift in terms of their structure, topology and operation. These major changes open new areas of research for capacity enhancement and energy optimization.

Existing macro cellular infrastructure reached its capacity limits and possibly will fall short of meeting forecasted traffic demand. In the past decades, research was motivated by the desire to enchance capacity without altering network topology. In these research efforts, planned cellular networks are optimized to meet peak traffic demand. Optimizations took place in spectral efficient wireless transmission techniques, such as orthogonal frequency-division multiple (OFDM), and multiple-input multiple-output (MIMO) techniques. Besides, sophisticated wireless resource management and mobility management schemes have also been put forward to ensure the quality of service (QoS) experienced users. Despite their usefulness, these enhancements will not be able to alleviate traffic growth problem.

Approximately 1000x traffic throughput will be needed in the next ten years according to 3GPP [1] and Qualcomm's 1000x data challenge [2]. One way to achieve such capacity is to increase frequency reuse by deploying small base stations (SBSs) [3, 4]. Besides the overall traffic increase due to the growing number of mobile users, the reasons for the deployment of small base stations lies in diverse QoS demands of mobile applications and dead zones (e.g. inside buildings, or in the subway). As a promising technique for future networks, small base stations expand the coverage of cellular networks, provide high bit rate, decrease the transmission power of user equipment, and increase the spectrum efficiency.

In communications networks, capacity and energy demand problems come hand in hand. Solving capacity challenge this way will bring energy-efficiency problems since small loads in small base stations will cause them to stay idle and waste power. To tackle this issue, it is necessary to develop energy-efficient techniques that not only meet capacity demands but also avoid wasting power in small cell networks. It is possible to design energy-saving methods by taking into account the hardware flexibility of small base stations, smart phone usage patterns, and delay tolerance of data applications.

Small base stations (aslo called femtocells, picocells, and microcells) are commonly seen as a solution to meet this growing data demand. Despite boosting network capacity, large-scale deployment of SBSs will inevitably bring some challenges:

- *Irregular topology*: In conventional cellular networks, cell patterns are roughly in the form of regular shapes such as hexagons, circles, and squares. Such patterns and assumption of symmetrical deployment may not hold for small base stations. Besides, in large scale deployment, the irregularity of SBS locations inevitably affect many performance metrics.

- *Interference management*: Conventional techniques such as frequency reuse, and sectoring for interference management in planned networks may not be applicable in small cell networks due to irregular SBS locations.

- *Resource scheduling*: Frequency resources needs to be shared between small base station and macrocells. One solution of frequency assignment is having dedicated frequency bands. However, this solution reduces spectral efficiency. Therefore, an efficient but challenging solution is to operate small cells in a co-channel band with existing macrocell.

- *Energy consumption*: Small base stations conserve much less energy when compared to macrocell. However, large scale deployments of small base stations still consume considerable amount of energy. In [5], operational costs of small base stations and macrocells are compared. It is shown that cost of one macro base station amounts to total operational cost of 450 small base statations. Therefore, in large scale deployments, energy-efficient mechanisms are still necessary to avoid both idle power consumption at low traffic periods and unnecessary signaling between user equipment (UE) and small cell.

Besides aforementioned technical challenges, enviromental concerns also stimulate the research efforts for energy-efficient networks. The growing demand for higher throughput, and peak-traffic based network planning have a high energy bill on global scale. The information and communication technologies (ICT) contributes 2% of global Greenhouse Gases (CO_2) emissions [6]. The amount of CO_2 due to information and communication technologies was 151 MtCO_2 in 2002. It also forecasted that the CO_2 emission will rise to 349 MtCO_2 in 2020, in which 51% of emissions are from the mobile networks [7].Telecommunications industry takes initiatives for climate change, eco-sustainability, and global energy management. For this reason, a number of collaborative projects, such as GreenTouch consortium [8], EARTH (Energy Aware Radio and Network Technologies) [9], GREENET [10], have been created. In conclusion, the growth of energy consumptions in ICT necessiates research on energy saving mechanisms for wireless networks.

1.1 Motivation and Basic Approaches

Design of energy efficient algorithms requires understanding of traffic dynamics, hardware capabilities of small base stations as well as data and voice service quality thresholds. There are many factors that affect mobile data usage. According to Jonsson et al. [11], monthly data traffic depends on the throughput limitations by telecomunications company, user tariff plans, device capabilities, display size and pixel quality of user device. Andone et al. [12] showed that age and gender affect smart phone usage. Zhao et al. [13] identified 382 different usage patterns among 106,762 Chinese smart phone users. Considering these findings, significant spatial and temporal variations can occur over the course of day.

Regardless of aforementioned statistics, existing cellular networks are statically planned and their design relies on the peak traffic load. Besides, many energy-reduction techniques in literature are based on average traffic demand. Marsan et al. [14] found an optimal time to switch off base stations assumuing average traffic intensity has trapezoidal pattern during day and night. Similarly, in [15] daily change in traffic intensity is modeled by cosine function. Although these approaches are successful at saving energy, they may not be able to adapt short term traffic fluctuations. Besides being good candidate for boosting capacity by increasing frequency reuse, small base stations have a low-power hardware which paves the way for flexible switch off/standby operation during the inactivity periods. Taking advantage of the fast on/off adaptation of small base stations, it is possible to design energy-saving mechanisms that are capable of adapting to both long term and short term traffic changes.

An important aspect of wireless mobile traffic is that in some cases, user equipment can tolerate modest delay before communication begins. First, we will try to determenine whether and when this delay can be utilized to save energy. Then, we will explain large scale effects of delay tolerance in terms of energy efficiency. Eyers and Schulzrinne [16] points out that %95 call set-up delay is below 6-11 seconds in VoIP with a critical threshold of 25 seconds [17]. According to Galletta et al. [18], response time of web site must be less 8 seconds in order to avoid leaving negative impression to users. Also, they conclude that response time should be less than 4 seconds to keep users interacting. In [19], 53% of mobile users leave sites if the response time exceeds 3 seconds. So, critical threshold for initial access delay for web response is somewhat tight when users are actively engaging with cell phone or tablet.

There is non-negligable *background traffic* in wireless mobile network. We define *background traffic* to be traffic between base station and user equipment when user is not actively using smart phone or tablet such as email, social media notifications. According to Meng et al. [20], between 33% and 40% of total traffic across the wireless interfaces of user equipment is background traffic. Huang et al. [21] measured traffic from 20 users and studied screen-off radio energy consumption. They found that 35.84 % of total traffic is background traffic corresponding to 58% of radio energy consumption. They proposed a delay tolerant access scheme that opportunisticly offloads background traffic to WiFi and saves energy. It is reasonable to assume that background traffic can tolerate larger delays compared to the traffic occuring when user is active, and its proportion in all data traffic during low load periods (e.g. night time) is higher than that of day time. Besides background traffic, there are other applications that make intermittent connections such as electronic meter readers, rental electric bikes, and scooters.

It is possible to turn small base stations off without distrupting active communications. In [22], Moon et al. showed that after losing connectivity due to mobility, some popular applications can still tolerate up to 5 minute delay, and seamlessly recover network connection. They also developed a socket API that conforms to TCP, and transparently handles network mobility when there is distruption of connection. In case small base stations are turned off to save energy, connection failures can be avoided, and delay tolerance of background traffic can be managed in an energy efficient manner.

Taking advantage of delay, not only a variety of user-cell association mechanisms but also energy-efficient schemes can be developed. If user equipment defers its access to the small base station, it may have the opportunity to find a closeby cell, which decreases its service time and increases spectral efficiency. Then, more small base stations can be turned off in the small cell networks compared to the case in which user equipment has no delay tolerance.

Conventional cell selection mechanisms are practical but often not optimal in terms of capacity and energy-efficiency. Cell selection based on the instantaneous signal strength may not be optimal [23] because signal strengh randomly changes due to fading and scattering. Associating user to the nearest cell is another common practical method, which may cause imbalance in the distribution of load. Besides, the algorithms for optimal user cell-association are computationally intensive and do not take into consideration of short-term traffic fluctuations. It is necessary to design practical access schemes for small cell networks which not only take advantage of user's initial access delay but also operate in a decentralized fashion.

In this dissertation energy efficiency of small cell networks are investigated under different access delay margins tolerated by the user equipment. Unlike complex resource allocation, and user-cell association strategies, the schemes introduced in the dissertation are kept rather practical in order to analytically track the large scale effect of delay tolerance.

1.2 Assumptions and Limitations

Throughout the dissertation, we made several assumptions regarding network topology, traffic, user mobility and localization. We will provide succinct information underlying these assumptions. In the following Chapters, we will also provide more detailed discussion regarding these assumptions.

We assumed that location distribuitons of small base stations and user equipments follow Poisson point process with parameters ρ_{c}, and ρ_{u} respectively. This assumption allows the analysis of important performance indicators such as coverage probability, throughput, and delay-energy tradeoffs, which will otherwise be intractable. We will give more detailed discussion regarding network topolgy in Section 2.4.1.3.

User equipment initiates a service request at random Poisson intervals with rate λ_u. Service time is either assumed to follow a distribution or computed by using Shannon's capacity and size of offloaded file. We clarify the modeling of service time in each chapter. For the sake of analytical tractability, we omit the case of periodic arrivals.

The scenarios considering diverse traffic patterns are omitted and left as future work. However, we briefly discussed possible ways our proposed algorithms can be extended when multiple traffic patterns are in play.

It is also assumed that location of user equipment can be obtained with reasonable accuracy by existing methods in literature [24, 25]. Having said that, for the user centric access schemes, we assumed that UE can play active role in initiating and terminating connection with SBS based on signal strength and distance to SBS.

Energy efficiency is not computed considering macrocell tier, we focused on energy efficiency of small cell network. It is assumed that user equipment gives the priority to small cell when it needs to receive service.

Frequency reuse is assumed throughout the dissertation. Performance of energy saving algorithms can be futher improved considering frequency allocation along with proposed schemes. We used Shannon's capacity formula to comput wireless link capacity.

For simulation scnearios, we aim to evaluate peformance of algorithms at low, and medium utilization levels. We choose the UE and SBS density so that utilization levels are met. Since the range of SBS can scale from 10 m to 1km as shown in Table 2.1, range of possible UE and SBS density is large.

1.3 Organization of the Dissertation

The scope of this dissertation is twofold: *i*) Development of cell selection strategies for user equipment tolerant of initial access delay, *ii*) designing energy-saving algorithms for small cells. The key design element of our access schemes and on/off algorithms is to combine initial access delay with the hardware flexibility of small cells.

This dissertation deals with the design and analysis of energy-saving algorithms for small cells. In Chapter 2, we provide a summary of the following subjects: *i*) metrics for energy-efficiency, *ii*) power consumption models for small cells, *iii*) evaluation of energy-efficient schemes for small cells.

In Chapter 3, we introduce a simple and practical delayed access scheme and analyze its effectiveness in small cell networks with on/off capability. We analyze a simple energy-saving model operating in a random fashion. On the network side, small cells turn on and off randomly while user equipment makes a decision to connect to the closest available cell within the delay budget. Also, the optimality conditions of this access decision will be discussed. Moreover, in terms of transmit power of user equipment, we will show the contrast between two small cell networks, wherein in the first scneario, a set of inactive cells are changing randomly and in the other scneario topology remains static as a hexagonal grid.

In Chapter 4, a new metric that measures traffic load to the cell is defined. Based on this metric, the traffic load distribution for a given small cell is obtained by a Gamma distribution approximation. Our numerical results show that the network throughput, and energy-efficiency can be improved considerably.

In Chapter 5, stochastic geometry tools are used to analyze delayed access scheme in random topology network. Distribution of coverage probability and average bitrate are derived. Optimal transmission range maximizing bitrate and coverage probability are discussed. Results are verified via simulations.

In Chapter 6, we offer a preliminary discussion about delay capacity tradeoff arising from the applications that both use large bandwidth and impose high computational loads.

Conclusion part summarizes energy-efficient schemes discussed throughout the dissertation, and points out possible venues to expand the research work.

Chapter 2: Background and Literature Review

Small base stations are expected to play key role in expanding the capacity of wireless network, and satisfying the growing traffic demand. With the advent of small base stations, the need for careful network planning and optimization is also increased. In this chapter, we introduce basic features and main challenges of small base stations with an emphasis on energy-efficiency concerns. We will explain why small base stations are deployed and challenges small cell deployment brings. We will also discuss energy-efficiency metrics, energy-consumption models, and comparison of energy-efficient schemes.

2.1 Types of Small Base Stations

A small base station is a low power access point equipped with radio frequency (RF) component. It can be both deployed indoors and outdoors. Advent of small cell started with the idea of frequency reuse indoors. Kinoshita et al. [26] suggested that if frequency channels are reused indoors with low power transmitters total required frequency bands can be decreased significantly. Then, in 1996, Silventoinen et al. proposed *home base station* [27] which is an indoor base station with small coverage that co-exists with macrocell provided that interference from macrocell is controlled properly. 3rd Generation Partnership Project (3GPP) has released a number specifications for small cells [28, 29, 30].

Small cells can operate in both licensed, unlicensed or shared spectrum [32]. We use *small cell*

Table 2.1: Small Cell Classification [31]

Cell type	Range	Deployment
Femtocell	10 m ~ 50m	Indoor
Picocell	100m ~ 300m	Indoor, Outdoor
Microcell	250m ~ 1km	Outdoor

and *small base station* interchangeably. Small cells have a range of 10 meters up to a kilometer. There are three kinds of small cells; femtocells, picocell, and microcells; listed in order of increasing range. These terms are not standardized, and may have overlapping usage. Table 2.1 gives a simple classification of small base stations based on range. Considering their on access control mechanisms, small cells can be classified as follows:

- *Open access*: Cell access is granted to any user equipment belonging to the operator that deployed the small cell. This type of cell is also called Open Subscriber Group(OSG).
- *Closed access*: Cell access is granted to only a subset of users by the small cell owner. This type of cell is called Closed Subscriber Group (CSG).
- *Hybrid access* Access priority is given to CSG user, and limited resources are allocated to non-CGS users.

We avoid focusing our study on single access mechanisms above. However, we envison that user equipment may communicate with multiple small base stations in large scale deployment. Small Cell Forum recently released a specification that enables an open, multivendor platform that eases densification for all stakeholders [33]. Besides, hardwave of future small cells is expected to support both licensed (LTE), and unlicensed bands (WiFi) [34]. Trend in the evolution of technology indicates that small cells will have many other functionalities. So, instead of considering only CSG type cells, we assumed that small cells are open access.

2.2 Metrics for Energy-Efficiency

In the literature, different energy-efficiency (EE) metrics (shown in Table 2.2) are proposed at the component, base station and network levels [35]. The component level, the ratio of power amplifier (PA) output power to input power known as ROI is used to denote the EE of the PA component, and MIPS/W (millions of instructions per second per watt) or MFLOPS/W (millions of floating point operations per second per watt) are used to calculate processing associated energy consumption. At the base station (BS) level, there are several of EE metrics to evaluate the energy efficiency. For trade-off between energy consumption and spectral efficiency (SE), bits per second per hertz per watt (bit/s/Hz/W) is defined. Also, bit times meter per second per hertz per watt (b·m)/s/Hz/W measures energy efficiency when energy consumption, transmission range of the base station, and spectral efficiency are all considered. At the network level, the EE metric is used to evaluate the obtained service relative to the consumed energy. The services include subscribers and the coverage range. Commonly, the ratio of number of subscribers served during the peak traffic hour to power consumption is used for urban environments, and the ratio of area covered by the network to the power consumption is used for rural areas. The consumed power per area unit (W/m^2) is used to evaluate the coverage EE.

Table 2.2: Energy Efficiency Metrics

EE Metrics	Level	Descriptions
MIPS/W or MFLOPS/W	Component	Used to calculate processing associated energy consumption
bit/s/Hz/W	Base station	For trade-off between energy consumption and spectral efficiency (SE)
(b·m)/s/Hz/W	Base station	Taking into consideration energy consumption, SE and the
W/m^2	Network	Used to evaluate the coverage EE

In large scale deployment, arhitecure of legacy base stations may not efficiently handle traffic growth. Conventional base stations have two main components: baseband unit (BBU), and remote radio head (RRU). Baseband unit carries out digital signal processing tasks. Remote radio head converts analog signal from digital baseband signal, then feeds it to the antenna. It also digitizes received RF signal. In dense networks, having dedicated BBU for each base station comes with operational cost such as cost of power and cooling. To realize cost-effective deployment, baseband units are aggregated in a pool [36], and remote radio heads are deployed at cell sites. Fiber links are used for communicaton between RRU and BBU.

Except its size, the function of small cell is the same as macrocell. Because of this similarity, the energy-efficiency (EE) metrics on the component level are suitable to both small cell and macrocell. As small cells are deployed with existing macrocell systems and use the same spectrum with the macrocell, it is reasonable to evaluate EE of macro base stations together with the small cell network. Besides, EE of small cell network needs to be analyzed considering the difference between the small cell-supported service and macrocell-provided service. It is also necessary to take the interference between small cell and the macrocell into account. For instance, the EE metric for femto-macro heterogeneous networks is defined [37] as,

$$EE = \frac{R_{Me} + \sum_{i=1}^{n} R_{He,i}}{P_{Me} + \sum_{i=1}^{n} P_{He,i}}, \tag{2.1}$$

where R_{M_e} denotes the average data rate provided by the macrocell with consumed power P_{Me}, $R_{He,i}$ denotes the average data rate provided by the i^{th} femtocell with consumed power $P_{He,i}$. (2.1) considers the service rate and consumed power in both the macrocell and femtocell. However, the calculation of the service rates of the macrocell and femtocells is a challenge because the macrocell and femtocells overlay with each other, and the service rates are related to the detailed management schemes, such as the resource allocation, interference cancellation, etc.

2.3 Power Consumption Models for Small Cells

In general, energy consumption of a wireless network is evaluated at two levels: operational energy consumption and embodied energy consumption. The operational energy that is defined as the amount of energy spent during a system's operational lifetime varies with different configurations such as load of the cell and RF power efficiency. It is generally accepted that most power consuming component of base station is RF power amplifier. So, in parallel with evolving communication technologies, there is also enourmous research and investment made every year to improve the efficiency of RF power amplifier [38]. The embodied energy is defined as the total primary energy consumed in the work of making a product. Power consumption models and detailed power consumption values for small cells are given in [39, 40]. Component wise power consumption levels are given in Table 2.3.

Table 2.3: Power Consumption of Femtocell [39]

Hardware component	Energy Consumption (W)
Microprocessor and associated memory	2.2
FPGA associated and associated memory	2.5
Other hardware components	2
RF transmitter	1
RF receiver	0.5
RF power amplifier	2

Power consumption depends not only on the hardware but also on the communication technology such as HSPA (high speed packet access), WiMAX, LTE. According to the model proposed by Deruyck et al. [41], femtocell comsumes 10.5W, 10.0W, 9.7W for WiMAX, LTE and HSPA respectively due to the difference in input power of antenna. So, typical femtocell consumes about 10W when fully active. Picocell and mircocell consumes 40W and 80W respectively [42] The embodied energy of a cell is assumed to have a similar value to a mobile terminal which is 162 MJ. For instance, assuming the lifetime of a femtocell is about 5 years, its embodied energy per sec is about 1 W [43].

There are several power models with varying complexity. The simplest one is on-off model. In this model a small cell is assumed to spend unit power in active mode, and zero power in off mode. The model is useful for theoretical analysis. Auer et al. [40] proposed linear power consumption model (2.2) This model is widely accepted and used with slight variations in the literature [44, 45, 46].

$$P_{\text{tot}} = \begin{cases} N_{tx}(P_0 + \Delta_{\text{c}} P_{tx}) & \text{if active mode} \\ N_{tx} P_s & \text{if sleep mode,} \end{cases} \tag{2.2}$$

where N_{tx} is number of tranceivers, P_0 is the static energy consumption excluding the tranceivers. Δ_c is the load dependent energy consumption coefficient, and P_{tx} is the transmit power of small cell. is the backhaul power consumption for the small cell. For example, for $N_{tx} = 2, P_0 = 4.8$

$P_{tx} = 0.05\,W$, $\Delta_{\mathrm{c}} = 8$, $P_s = 2.9W$, power consumption of LTE femtocell $10.4\,W$, and 5.8 in fully active and sleep mode respectively. Also, if traffic load is zero ($P_{tx} = 0$), power consumption of femtocell is $9.6\,W$. Relationship between power consumption and traffic load varies with base station type. Since the coverage range of femtocell is small, P_{tx} is relatively small. So, traffic load has less impact on overall power consumption. For marco cell, P_{tx} is large; hence, traffic load has a significant impact on total power consumption.

Power consumption at different hardware components brings extra challenges and tradeoffs in designing energy efficient algorithms. The linear power model is also useful for analysis and it is more accurate than the on-off model. However, it may not reflect power consumption in sleep mode accurately. Power savings in sleep mode depends on *depth*. As switching off more components, energy savings proportionally increase. However, increase in energy savings comes with a cost of reactivation delay. For femtocell, if RF unit is turned off only, energy savings are about 40-50%. ([39, 40]). Vereecken et al. [47] proposed multiple sleep modes in decreasing power consumption and increasing reactivation delay. They suggested on, stand-by, sleep and offline with the information of corresponding power consumption and reactivation delays. Liu et al. have measured power consumptions based on multiple operation modes and boot-up delays. [42].

In this thesis, we assume small cell can be in *sleep*, *idle*, and *active* modes. In active mode, we assume small cell has active communication with a user equipment. In idle mode, small cell is active but does serve any user. We have slightly different approaches for sleep modes. In Chapter 3, we assume that radio part of small cell is turned off in sleep mode and there is no boot-up delay. In Chapter 4, there are multiple sleep modes with associated boot-up delays (see Table 4.1).

- *Active*: The SBS is actively engaging in transmission with full transmit power.
- *Idle*: The SBS is active but not serving any user.
- *Sleep*: The SBS is in a sleep with only necessary hardware parts. Power consumption varies depending on the components switched off. Sleep modes are clarified in chapter.

- *Off*: The SBS is completely offline.

Last but not least, it is important to design an energy efficient scheme in which activation delay does not adversely affect quality of service for both rapid and long term traffic fluctuations. To this end, it is important to choose both length of the sleep period, and the hardware components to deactivate in these sleep periods. To tackle this problem, it is proposed in [48] to turn off only the transceiver part of base station for short sleep periods and shut down the entire system during long sleep periods. Finally, average power cunsumption of small cell is computed by taking the product power consumption levels and proportions of times that small cell is idle, active or in sleep mode. We summarize small base station operation modes as follows

2.4 Evaluation of Energy-Efficient Schemes for Small Cells

In this section, we review energy-efficient schemes for small cell networks. We first classify them based on their design objectives, and discuss briefly their advantages and limitations, and consider system performance measures.

2.4.1 Classification of Energy-Efficient Schemes for SCN

We review energy efficient schems in several categories: power control schemes, dynamic idle/sleep schemes. We also discuss deployment strategies. Power control schemes are mainly designed for interference management. However, they still benefit energy efficiency as the antenna transmit power is decreased.

2.4.1.1 Power Control Schemes

Although the main purpose of power control is to cancel or coordinate the interference (i.e. decreasing the interference level in the vicinity of a small cell), power control schemes obviously decrease the transmission power of small cell, and make them energy efficient. According to the contexts on which power control schemes are based, we roughly classified them into receiving power based, traffic load based, frame-utilization based, and global energy based schemes.

Receiving power: Controlling transmit power of small cell based on the strength of received power from user equipment is standardized by 3GPP. Standardization on the power control schemes vary with the type of small cells. Therefore, without delving into much detail, we briefly discuss the femtocells' down link (DL) inter-cell interference cancellation (ICIC) strategy. In [37], two autonomous power control (APC) schemes have been proposed. Small cells that have closed access or hybrid accress policy bring the risk of interfering user equipments communicating with macrocell. A non-subscribed user in the vicinity of closed access cell (CSG cell) has to communicate with macrocell; thus, it may receive strong interference because of shared spectrum. In this case, the user equipment may become *victim* of high interference from CSG cell. Similarly, macrocell can be exposed to interference from a closeby small cell user. To avoid the risk of interference from closed access cell, a femtocell detects the existence of victim users interfered by the femtocell through measuring up link received power from a victim user; if the measured power is larger than a pre-determined threshold, the femtocell believes that there at least exists a potential victim. After that, 3GPP recommends several schemes to adjust the transmission power of the femtocell. One method to avoid the interference is through reported reference signal received power (RSRP), and reference signal received quality (RSRQ) measurements from subscribed user. Based on the measurements small cell optimizes its transmit power so that predetermined thresholds for RSRQ and RSRP are met. Similarly, small cell can decode RSRP from macrocell user. If RSRP from the nearest macrocell user is poor, the transmit power of small cell set to lower value in order to mitigate the downlink interference with macrocell user. Another way is to make use of GPS signal. If small cell detects GPS signal well, it indicates that it is deployed outdoors, and there is a risk of interfering macrocell users. So, transmit power of small cell should be small. Conversely, if the GPS detection is poor, small cell is likely to be deployed indoors, or in a dead zone to extend coverage. In that case, transmit power of small cell can be large. Suggested techniques are validated by simulations. To enhance the feasibility and performance of the recommended schemes, in [49], a strategy for femtocell to appropriately determine the self-configured threshold is used to autonomously trigger DL ICIC. Also, effective strategy to autonomously estimate the existence of indoor victim users is proposed and validated by simulation.

Traffic load: Based on geometric Voronoi diagrams, [50] Shin and Choi proposed a dynamic power control algorithm according to data traffic estimation, which achieves good indoor coverage and load balance in experiments.

Frame utilization: In [51], Mach and Becvar suggested frame-utilization based power control scheme. In this scheme, femtocell decreases its transmit power, and uses more resources blocks in LTE frame. To avoid interference, femtocell transmits with low power. This will cause number of decoded bits per symbol to be lower. In other words, femtoell adopts lower modulation and coding scheme (MCS) in expense of using larger frequency and time resources in LTE frame while maintaining the same QoS for user equipment. This way the interference range and average transmit power of femtocell are decreased.

Global energy approach: In [52], Chen et al. suggested a method solving an potimization problem which is defined as maximizing aggregate network throghput in small cell network for given bandwidth. Defining the problem as maximizing the sumrates over all users may seem a solution. However, in practice, solution offers very low throughputs to some users, which is unfair. A more reasonable appraoch is minimizing inverse of signal to interference plus noise ratio (SINR), which can also be interpreted as minimising delay experienced per bit. To be mor specific, small cell needs to be assigned a transmit power; having discrete values between 0 and P_{max}, a channel among a set of channels. Possible channel and power assigments have a large space. To solve the problem, global energy function which is defined as 1/SINR and its sum over all users is minimized by the Gibbs sampling technique. Gibbs sampling defines a joint distribution function taking channel and power assigment values as parameters. Using this distribution, joint distribution is obtained. Parameter values are updated iteratively until they converge. Efficiency of global energy and Gibbs sampling solution is validated by simulation.

Although there are many publications concerning power control schemes for femtocell or femto-macro heterogeneous networks, most of them focused on interference cancellation/coordination but not energy efficiency [53, 54, 55]. When more and more femtocells are deployed, coverage region of femtocells will overlap. It is necessary for neighboring femtocells to collaborate on power control to cancel the interference and improve energy efficiency.

2.4.1.2 Dynamic On/Off Schemes

Small cells and macrocell differ in the traffic pattern. The coverage region of a small cell is smaller than that of macrocell. A small cell commonly serves a few users, especially when it is deployed indoors (i.e. femtocell). So, it is likely to be idle as there are no users requring service during low traffic periods, On the contrary, macrocell, even in low traffic, is likely to serve a few active users. Due to this difference, energy efficient schemes are tailored to suit traffic dynamics and hardware capabilities of small cells. We review related litureture below.

Sniffer based: In [39], the authors proposed two dynamic idle mode schemes, “Idle Mode Based on Noise Rise” and “Reducing Pilot Power When Idle”, which allow the femtocell’s transmitter and associated processing to be switched off completely when the femtocell does not need to support an active connection. System level simulations were run to demonstrate the effectiveness of the schemes. Their efficiency is still unclear in open-access femtocells since if a user is in the vicinity of multiple femtocells, the elevation of noise level might trigger multiple wake-ups unless a central controller is used.

Traffic load based: In [56], a sleep mode scheme is proposed for a single femtocell. In this scheme, a femtocell saves energy by aligning the listening windows of multiple mobile users. By doing this, the femtocell sends data to mobile users at the same transmit intervals. Minimum and maximum sleep intervals are defined, and adaptively changed based on traffic. Simulation results indicate that sleep ratio can be increased by 20% by the proposed scheme while maintaining same QoS level.

Localization: Using Markov Decision Processes (MDPs), Saker et al. [57] proposes optimal sleep/wake up schemes for macro-femto heterogeneous networks, in which femtocells work in open access mode and can offload traffic from the macrocell. The proposed sheme suggests that when the macro-femto heterogeneous network is not highly loaded, macro base station can handle the traffic itself, and femtocells are switched off. As the load increases, one or more femtocells are selected to be switched on. Simulation results indicate about 10% reduction in average consumption of femtocell.

Active connection based: In [58], Vereecken et al. offered heuristic on/off scheme that makes use of overlapping coverage regions of femtocells. It is assumed that topology of small cell network, locations of users, and their bit rate demand are known. Initially, user and small cell adjancency matrix is generated based communication range of small cells. The matrix gives the information about the number of small cells that can serve a certain user. This information is used to determine user's alternative connections to its currently associated cell. If user is in range of only one cell, the small cell that it is currently associated remains on. Then, small cells that serve the most users are turned on until such that each user has at least one cell in its communication range. The paper formulates integer program that minimizes number of on base stations. This heuristic is shown to be close to the optimal solution. Intended sleep durations are much longer than that of [56].

QoS constrained: In [59], the tradeoff between the energy-efficiency and the delay has been investigated using the metric of effective capacity [60], which characterizes QoS in terms of bits successfully communicated as a function of time with high probability. The problem is formulated as finding the optimal sleep and idle periods so as to maximize the energy efficiency subject to the effective capacity constraint, and solved by using genetic algorithms. Simulation results indicate that energy efficiency (bits/joule) can be doubled through sleep modes.

The main focus of power control schemes is to manage interference, improve throughput and conserve energy consumption while the radio frequency (RF) and hardware of small base stations are always on; accordingly, the energy efficiency of power control schemes is limited. In contrast, on/off schemes focus on designing various approaches to turn on/off the RF and some hardware parts of small base stations, which could potentially save more energy but cause longer boot-up delays than power control schemes. On the other hand power control schemes and idle/sleeping schemes are complementary. Taking into consideration of the 3GPP standardization efforts [61]

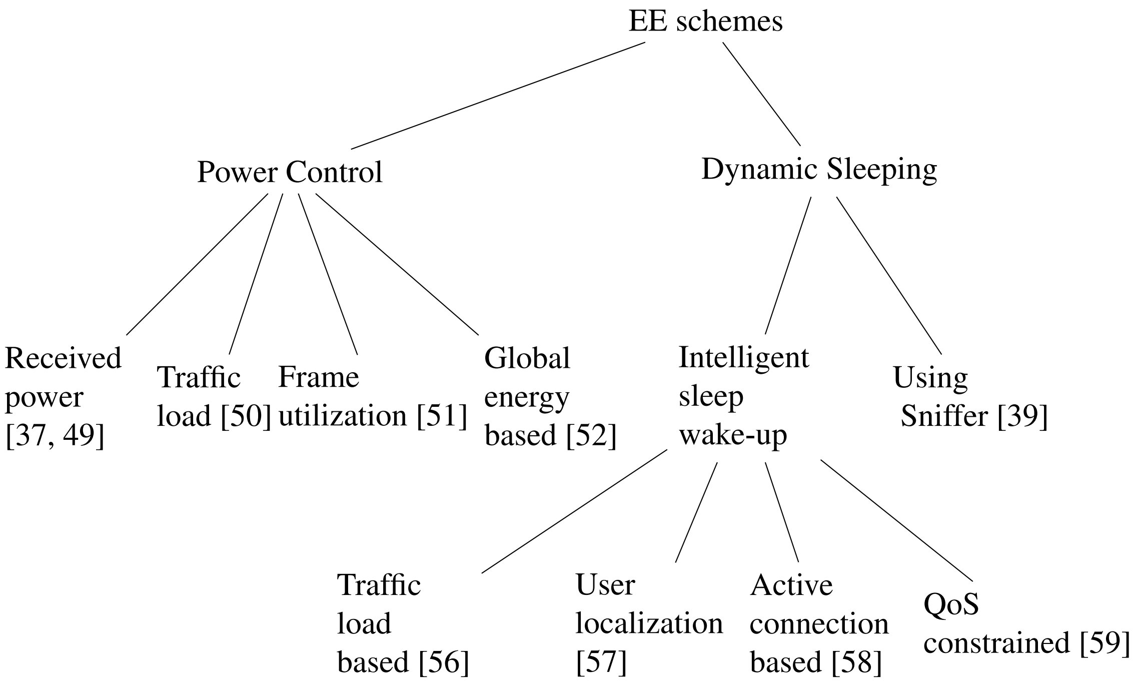


Figure 2.1: Classfication of energy-efficient schemes of small cell networks

for energy efficiency, both types of energy conservation schemes can be run Long Term Evolution (LTE) systems.

2.4.1.3 Deployment Strategies

Small cells are designed to extend the coverage range of cellular networks and support higher data rates in indoor, where the wireless signal is deeply deteriorated due to shadowing and multipath fading. Or, they can be deployed in outdoors, where traffic demand can be high such as Stadiums, and crowded streets. Figure 2.2 shows a general deployment of femtocells in indoor environments. We will first give an overview of relevant literature about small cell deployment methods. Then, we will discuss our assumptions regarding small cell deployments.

The relationship between energy efficiency and deployment of small cells is investigated in [63], where the energy-efficiency of downlink in marco-small cell heterogeneous networks is analyzied. The analysis shows that there exists a ratio of optimal small cell-macro cell density that maximizes the overall energy efficiency of heterogeneous networks. Considering the co-channel

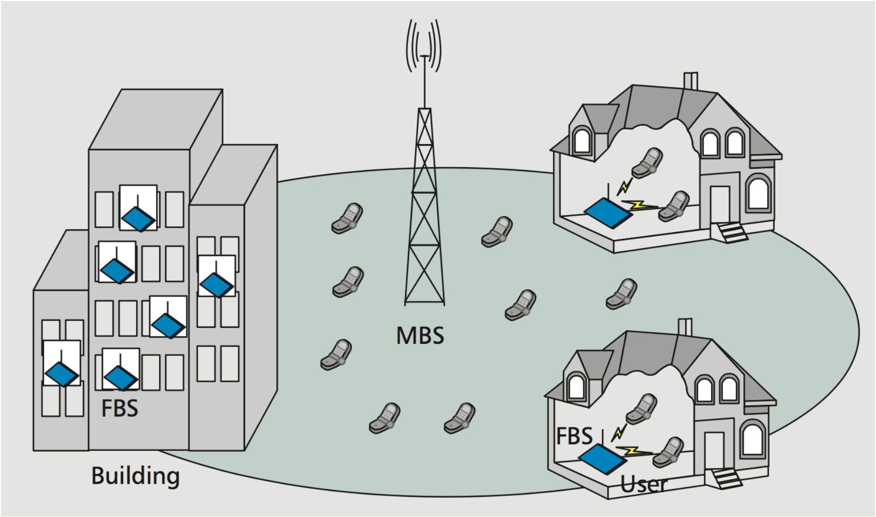


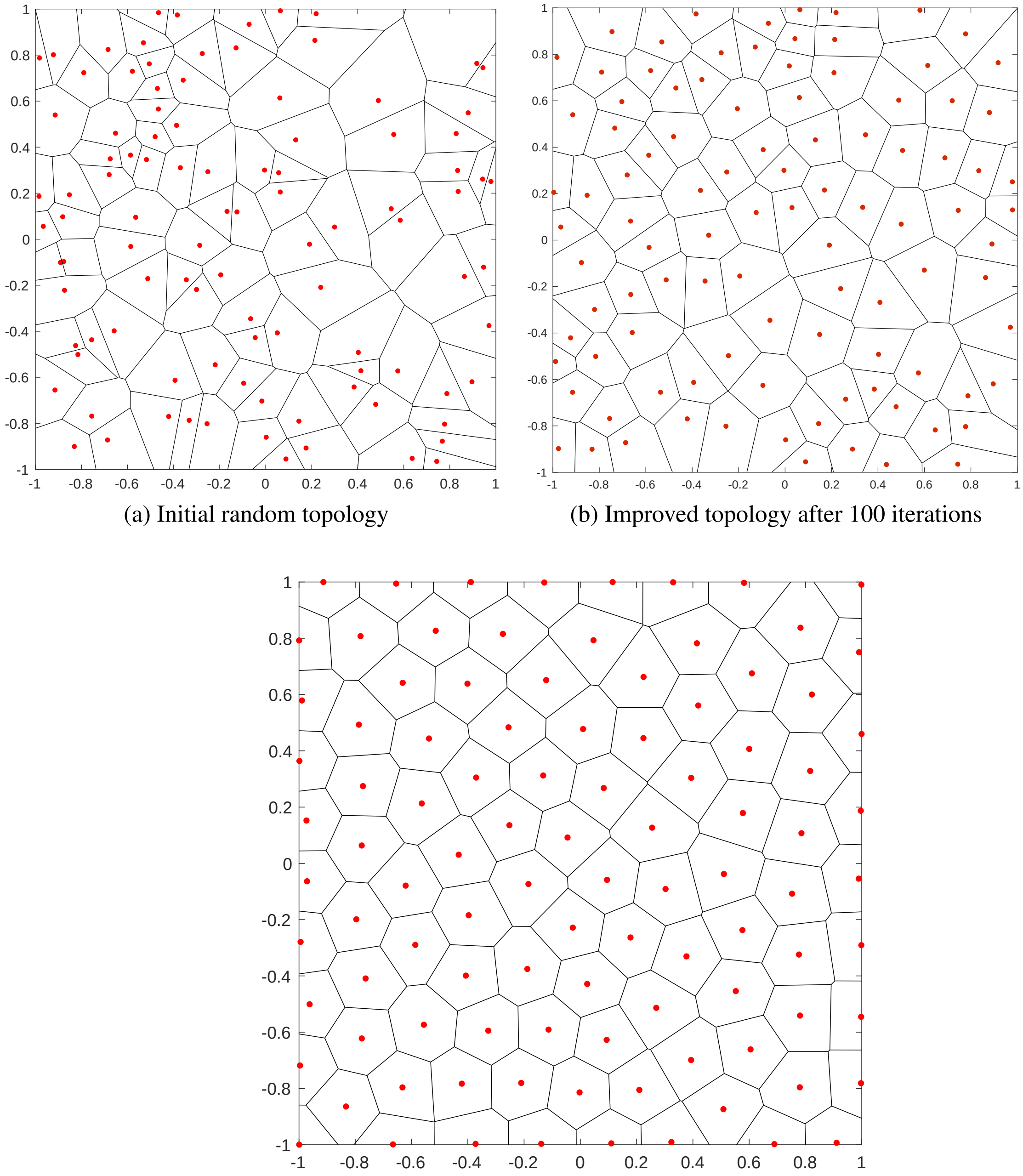
Figure 2.2: Femtocells deployed in the indoor environment [62]

interference and varying traffic demand, [64] evaluates the energy efficiency for heterogeneous OFDM-based mobile networks, and shows that increasing deployment density through additional small base stations may maximize a network's energy efficiency when the traffic demand is sufficiently large.

Although the assumption of random deployment of femtocells is somewhat reasonable, it is still questionable whether this deployment causes inefficiencies in QoS and power consumption. Inefficiencies in terms of high interference may also arise when small cells are deployed arbitrarily close. Recent work [65], compares average throughput and power consumption when small cells are randomnly placed with the strategy that places femtocells intelligently based on the position of other femtocells in the building. With and without outdoor interference sources, optimal positions to deploy the femtocell inside a room are derived. In case there are multiple femtocells in the building, a reasonable heuristics is used for the deployment strategy.

In practice, location of base stations are carefully planned by the operators taking into consideration of many factors such as traffic demand, line of sight, and enviromental concerns. So, topologies of real cellular networks are neither perfect hexagons nor completely random [66].

To understand pros and cons of grid and random topologies, we made a simple simulation that shows the transition from random topology to the perfect hexagons. In area A N small cells are



(a) Initial random topology

(b) Improved topology after 100 iterations

(c) Near hexagonal topology

Figure 2.3: Simulation example with $N = 100$ cells showing how completely random topology turns into *near* hexagons by iteratively changing inter-cell distance. Simulation starts with a random topology as in (a). After 100 steps, topology forms as in (b) For large number of iterations, topology is almost hexagonal as in (c)

placed randomly. Then, this random topology is *improved* in terms of intercell ditance. A new small cell is placed randomly in area A. Then, pair of closest small cell is found. The one, whose distance to second nearest base station is smaller is removed. This way, at each iteration, interell-distance is improved. Figure 2.3 shows the transition of random topology into *almost* hexagonal topology.

At each iteration, variance of inter-cell distance becomes smaller and minimum distance between closest pair becomes larger. Improvement in terms of inter-cell distance happens in two ways. With high probability, a new added cell is not removed in random topology since it is more likely that new cell is dropped at a sparse region than dense region.Thus, topology improves by densification of sparse regions, and sparsification of dense regions. As the number of iterations increase, voronoi cells with random size form into *near* hexagons.

Major disadvantage of random topology is that locations of small cells can be arbitrarily close. In grid model, effect of co-channel interfernece is minimal; bringing better QoS in terms of throughput. However, grid model is too idealized for large scale deployments because distance between two cells may vary depending on physical constraints and demographics. Interference is analyzed for a single user at fixed location in worst case scnenario [67]. Hard core point processes (HCPP) ensure a minimum hard core seperation distance between any pair, but it is difficult to analyze interference, delay-energy tradeoffs by using the complex expressions for the distance distributions of HCPP and regular polygons. [68, 69].

Despite the aferementioned disadvantages, assumption of random topology is adopted in many studies [70, 71, 72]. The reason is that the assumption of random distribution allows tractable analysis of key performance measures such as SINR in large scale networks. Thefore, to obtain energy-delay tradeoffs in dense deployments, we assume that small cells follow Homogenous Poisson Process (HPPP). Besides, despite the fact that cellular networks are designed to meet peak load, there will be significant imbalance between available cell and traffic demand at off-peak hours due to the daily population dynamics [73] in urban areas. We assume that random topology reasonably represents such imbalance, and can be improved with energy saving schemes that considers location of users, and cells and their communication range.

2.5 Conclusion

Energy efficiency in small cell networks is becoming more and more important with their deployment at dense areas. In this chapter, we give a brief survey on this issue in terms of energy efficiency metrics, energy consumption models, deployments, and energy saving schemes. Our review shows that there are many interesting issues on the energy efficiency of small cell networks to be investigated. We conclude that in dense deployments, it is crucial to have energy saving schemes not only for small cells but also Wi-Fi access points (AP).Considering the dense deployment of Wi-Fi APs with large overlapping coverage, and the design trends [34] that small cells will operate both in licensed and unlicensed bands, Wi-Fi APs have similar inefficiencies in terms of energy waste at low traffic load (e.g. no use night times in office buildings). We belive that our enegy saving methods that we discuss in following chapters also apply to Wi-Fi networks.

Chapter 3: A Random On/Off Strategy for Saving Energy in Small Cell Networks

New mathematical models are needed to gain insight into operation of small cells in large scale networks due to the shift from planned hexagonal grids to irregular deployments. There is rich literature about the analysis of signal to interference plus noise ratio (SINR) distributions, and achievable bit-rates for randomly distributed networks [74, 75]. However, little guidance is available for large scale behavior of energy efficent operation of small cells. From this respect, we introduce a baseline energy saving model. In this model, small cell has several operation modes including energy saving mode. Then, we analyze its large scale behavior in dense deployment. The key role of the analysis functions as a proof of concept of possible energy savings via a delayed access strategy. In the following, we give detailed description of the model, and analyze the model. Then, we discuss the simulation results, and finally give directions for more advanced models. This chapter is an extended version of work published in [76].

The rest of this chapter explains baseline models for cell selection, and analysis of on/off schemes, and verification of results via simulation. After describing the model, we analyze the probability distributions of small cell's operation modes. On the user side, a simple delayed access scheme is introduced. This scheme helps user equipment decrease average distance to the small cell it communicates. Regarding this delayed access scheme several peformance measures are also analyzed. We analyzed access probability distributions within predefined distance and range is analyzed. Also, we derive distributions of transmit distance, and an optimal transmit range that minimizes the average transmit distance. We discuss the bounds on transmit power gains by comparing transmit distances with and without access delay. Finally, analytical results are verified through simulation experiments.

3.1 Random On/Off Model

In this part, the model for random on/off strategy is explained. The model has two main components: energy-efficiency operation of small cell, and cell selection strategy of user equipment. The model is called *random* due to probabilistic nature of time of switch-off decision and sleeping period. Each small cell has a simple energy-efficiency operation. Operational modes of the cells are *sleeping*, *idle*, and *active*. In the sleeping mode, antenna and maybe some baseband processing units are switched off. Power amplifier, and FPGA is on. Thefore, small cell can switch from sleep mode to idle mode quickly. In the idle mode, cell is ready to give service but not actively serving any user. In active mode, small cell can give service up to C users simultaneously. Capacity C depends on small cell type. For example, off the shelf femtocell can serve 4 to 6 users at a time [77, 78]. For analytical tractability, we assume single sleep mode is employed and omit multi-level sleep modes. State transition diagram is shown in Figure 3.1. Having said that, model can be extended to energy a saving policy handling both short and long sleep periods by conditioning on sleep time, and find the probabilities of being in long and short sleeps. Average power consumption at long and short sleep periods can be computed.

A user equipment (UE) can delay its access in order to connect a closer small cell as shown in Figure 3.2. During UE's waiting, a closeby small cell can wake up and become available. This enables the UE and small cell to decrease their antenna transmit powers on uplink and downlink, and improves overall network energy efficiency. The longer a UE can wait, the more likely a closer small cell will become available. In other words, a UE improves network energy efficiency and transmission power at the expense of delay. We show that the transmit power of the UE and small cell is greatly reduced when compared with a system that turns off a fixed set of small cells. Also, we show that the average power consumption of a small cell can be decreased, while providing a bounded transmit distance with a high probability.

UE has two access modes, namely *rapid* and *delayed access*. In the rapid access mode, a UE may be connected to a further small cell if the nearest small cell is unavailable. We define the tolerable delay w_{t} as the maximum time that a UE can wait before a connection begins. Threshold distance R_{th} is defined as the distance within which UE is willing to initiate a connection with a small cell without waiting.

We propose a second mode referred to as the delayed access mode, if a UE can defer its access to small cell. In this mode, when a UE has a service request, it immediately initiates a connection with the small cell provided that there is an available small cell within R_{th} distance. If not, the UE waits for w_{t} time units. During this time, if an small cell within R_{th} becomes available, the UE accesses that small cell. If no small cell becomes available within a radius of R_{th} before tolerable delay expires, UE connects to closest available small cell in the network. In this case, it is clear that the available small cell is outside the threshold distance. Operation of the delayed access scheme is summarized in Figure 3.3.

Distance based schemes are proposed in many studies. Fanghänel et. al [79] proposed a channel assigment algorithm based on Euclidean distance. Likewise, decision of user cell assocation is made based on distance in [58]. Delayed access scheme assumes that UE is able to decide whether or not cell is within its communication range, R_{th}. This decision can be made by measuring received signal strength indication RSSI level.

Regarding the access strategy, rapid scheme is considered baseline and compared with delayed access. From energy saving perspective, random on/off is considered a baseline and it is compared with static cell topologies.

3.2 Analysis of State Probabilities and Approximate Access Delay Distribution

In this part, we derive the state probabilities of a small cell, and access delay distributions. Our analysis encompasses the scenario where a small cell can serve multiple UEs simultaneously. Theoretically, C can be arbitrarily large, and the model can be easily extended to $C = \infty$ case. We give further discussion about selection of C in Section 3.2.1. Arrival rate λ_{u} represents aggregate traffic load from user equipment. In our model, small cell has three operation modes, namely, *active*, *idle* and *sleep* modes as shown in Figure 3.1. Each UE requests a service with the rate of

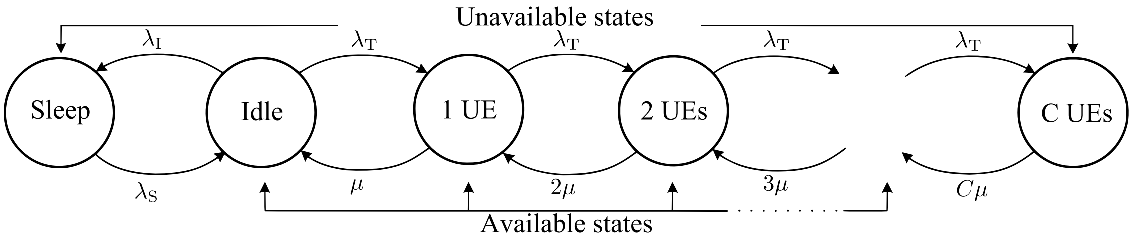


Figure 3.1: State transition diagram of a small cell

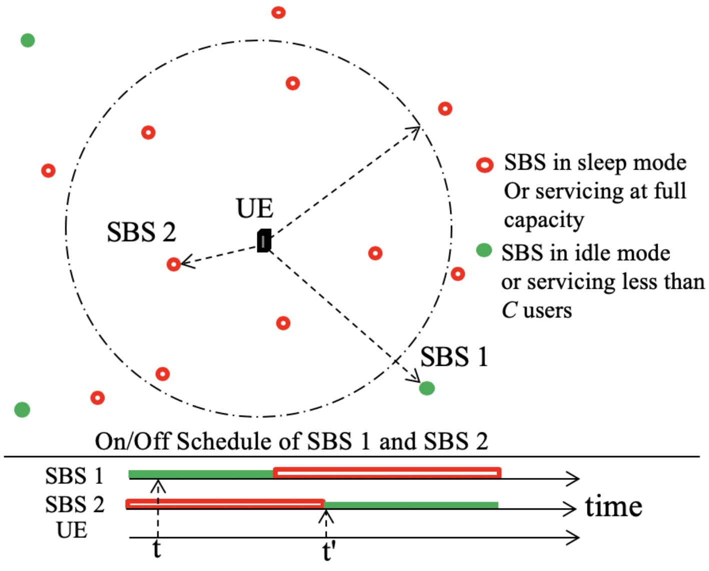


Figure 3.2: Illustration of transmit distance in delayed access scheme. The UE has a service request at time t. If the UE connects to small cell 1 immediately at time t, transmit distance will be R_{out}. If the UE delays its access and connects to small cell 2 at t′, transmit distance will be R_{in}.

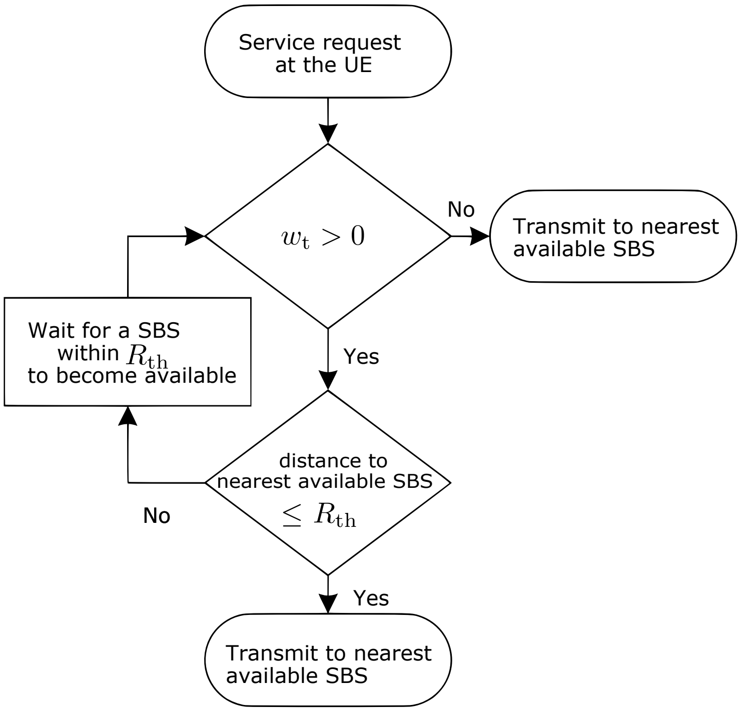

Service request at the UE
$w_{\mathrm{t}} > 0$
No
Transmit to nearest available SBS
Yes
Wait for a SBS within R_{th} to become available
distance to nearest available SBS $\leq R_{\mathrm{th}}$
No
Yes
Transmit to nearest available SBS

Figure 3.3: Delayed access strategy planned by user equipment

λ_{u}, and the service time is exponential with mean $1/\mu$. After small cell completes its service, it moves to the idle mode. Duration of the idle mode is exponential with λ_{I}. If a small cell does not receive a service request in the idle mode, it moves to the sleep mode. Duration of the sleep mode is also exponential with λ_{S}. For given traffic, UE density and small cell density, effective traffic load is $1/\lambda_{\mathrm{T}}$, which represents the mean time between two service requests to a small cell. λ_{T} is not an input parameter, but an internal parameter to be calculated. In other words, λ_{T} represents aggregate traffic load to the single SBS from UE's within its range. So, λ_{T} depends on user density, cell density, arrival rate as well as service rate.

In practice, there are various traffic patterns with different QoS characteristics. For example, regarding the arrival rate, traffic can be either periodic or aperiodic. It can be voice, data, or video traffic. It may or may not require guaranteed bit rate. Combining different traffic patterns in a single Makrov Chain is analytically intractable. Thefore, we prefered to use rather simple traffic model and discuss its strengths and weaknesses. We consider user equipment generates a service request at exponential intervals. Service request can be delayed up to w_t seconds, and service time is exponential. The traffic can be a background traffic which is rather delay tolerant as discussed in 1.1. Or, it can also be on-demand video traffic. So, user equipments downloads video segments from server periodically, and fills up its buffer in its storage unit. When buffer level is below its low threshold, user equipment downloads more video segments until its store unit is full. Depending on how fast buffer is emptied, user equipment knows the deadline to start downloading next video segments. Thefore, on demand video traffic is also delay tolerant. Although we made assumption of Poisson arrivals and exponential service times, we believe our model can capture the delay and energy efficiency tradeoffs.

Probability of sleep, idle, and active modes depend on the traffic generated by UEs, and density of UEs and small cells. As traffic load increases, sleep mode probability decreases. Operation of small cell is modelled as a Markov chain with three states.

The state-space is $\Omega = \{S, I, A_i\} | \ 1 \leq i \leq C$ where S, I, A are the events representing sleep, idle active modes; similarly, Π_{A}, Π_{S} and Π_{I} are active, sleep and idle probabilities respectively. The balance equations of the Markov chain [80] in Figure 3.1, can be written as

$$\Pi_{\mathrm{S}} = \Pi_{\mathrm{I}} \frac{\lambda_{\mathrm{I}}}{\lambda_{\mathrm{S}}}, \tag{3.1}$$

$$\Pi_{\mathrm{A_i}} = \Pi_{\mathrm{I}} \frac{(\lambda_{\mathrm{T}}/\mu)^i}{i!}, \tag{3.2}$$

$$1 = \Pi_{\mathrm{S}} + \Pi_{\mathrm{I}} + \sum_{i=1}^{C} \Pi_{\mathrm{A_i}} \tag{3.3}$$

The set of equations in (3.1), (3.2),(3.3) can be solved as a function of λ_{T}. Thus, the only unknown in our Markov model is λ_{T}. To solve λ_{T}, we consider equilibrium condition, where rate of service demand from UEs in the network is equal to the aggregate service rate. Then,

$$\begin{aligned} \rho_u \lambda_u &= \rho_{\mathrm{c}} (1 - \Pi_{\mathrm{S}} - \Pi_{\mathrm{I}}) \sum_{j=1}^{C} j \mu \Pi_{\mathrm{A_j}} && (3.4) \\ &= \rho_{\mathrm{c}} (1 - \Pi_{\mathrm{S}} - \Pi_{\mathrm{I}}) \mu \Pi_{\mathrm{I}} \sum_{j=1}^{C} j \frac{[\frac{\lambda_{\mathrm{T}}}{\mu}]^j}{i!} && (3.5) \end{aligned}$$

Left hand side of Eq. (3.5) is rate of demand from UEs per unit area. In other words, the product of density and arrival rate gives a measure of traffic in terms of number of arrivals per m^2. This traffic is met by the service rate on the right hand side. We have $(1 - \Pi_{\mathrm{S}} - \Pi_{\mathrm{I}})$ because only active cells gives service. Finally, plugging the average number of served connections by active cells using active state probabilities, equilibrium condition is satisfied. For $C = 1$, Eq. (3.5) deduces to following expression:

$$\Pi_{\mathrm{I}} = \left[1 + \frac{\lambda_{\mathrm{I}}}{\lambda_{\mathrm{S}}} + \frac{\lambda_{\mathrm{T}}}{\mu}\right]^{-1}, \tag{3.6}$$

$$\Pi_{\mathrm{A}} = \Pi_{\mathrm{I}} \frac{\lambda_{\mathrm{T}}}{\mu}, \tag{3.7}$$

$$\Pi_{\mathrm{S}} = \Pi_{\mathrm{I}} \frac{\lambda_{\mathrm{I}}}{\lambda_{\mathrm{S}}}. \tag{3.8}$$

To find λ_{T}, we consider the average number of UEs for which a small cell provides service. Considering the fact that only active cell can give service to a UE, and balance of arrival and departure rates in equilibrium, we have

$$\rho_{\mathrm{u}} \lambda_{\mathrm{u}} = \rho_{\mathrm{c}} \Pi_{\mathrm{A}} \mu. \tag{3.9}$$

Plugging Eq. (3.7) into Eq. (3.9), we obtain

$$\lambda_{\mathrm{T}} = \frac{\rho_{\mathrm{u}} \lambda_{\mathrm{u}}}{\rho_{\mathrm{c}} \Pi_{\mathrm{I}}}. \tag{3.10}$$

Inserting (3.6) into (3.10), we finally have

$$\lambda_{\mathrm{T}} = \frac{\mu \lambda_{\mathrm{u}} \rho_{\mathrm{u}} \left(1 + \frac{\lambda_{\mathrm{I}}}{\lambda_{\mathrm{S}}}\right)}{\rho_{\mathrm{c}} \mu - \rho_{\mathrm{u}} \lambda_{\mathrm{u}}}. \tag{3.11}$$

Note that $\rho_c\mu$ is a measure of the service that is processed in unit area, and $\rho_u\lambda_u$ is the service request from UEs in unit area. For network stability, we need $C\rho_c\mu - \rho_u\lambda_u > 0$, and network utilization is given by $\frac{\rho_u\lambda_u}{\rho_c C\mu}$. Note that network utilization does not depend on λ_S. Traffic demand is satisfied if stability condition $\frac{\rho_u\lambda_u}{\rho_c C\mu} < 1$ holds.

We can derive the access delay distribution using the small cell's operational state distributions that are provided in equations (3.6), (3.7). Access delay distribution is defined as the probability that a small cell within R_{th} is available in w_t time. The access delay not only depends on the state of the small cell, but it also depends on the number of waiting users for non-zero w_t. Here, we give an approximate distribution for the access probability assuming that the competition between UEs waiting for the same small cell is rare. Details regarding this approximation are given in the following section.

Let β_0 and β_{w_t} denote the probabilities that an small cell is idle immediately or will be idle within a duration of w_t, respectively. Then, using Eq. (3.7), for $C = 1$ we have

$$\begin{aligned} \beta_0 &\approx 1 - \Pi_S - \Pi_A, \\ \beta_{w_t} &\approx 1 - \Pi_S e^{-w_t\lambda_S} - \Pi_A e^{-w_t\mu}. \end{aligned} \tag{3.12}$$

By Poisson thinning property, available small cell density is $\beta_{w_t}\rho_c$. Then, using (3.12), with a delay budget of w_t, the joint probability to access an small within a radius R around a UE is given by

$$\mathbb{P}(R < r, t < w_t) = P_{in}(R, w_t) \approx 1 - e^{-\beta_{w_t}\rho_c\pi R^2}, \tag{3.13}$$

which gives the access-delay distribution. We will use Eq. 3.13 in Section 3.4.3 for the optimization of transmission range.

3.2.1 Number of Simultaneously Served UEs

It is useful to discuss the parameter C affects behavior of small cell. First, there is an upper limit on the number of users that small can be served simulatenously [77, 78]. Depending on the traffic load, UEs that have active connection with small cell also change. We discussed that it is possible to obtain state probabilities using balance equations in (3.1, 3.2, 3.1), and (3.5). We assumed that service time of active user have exponential distribution with parameter μ. In practice, if there are multiple active users that are connected to the same small cell, service of each user depends on how small cell allocates its frequency resources. At this point, it is necessary to make reasonable assumption on C so that energy-delay tradeoff can be characterized successfully. In our study, we consider that small cell network is underutilized. Consider a femtocell with three UEs. Each UE has arrival rate $\lambda_u = 1/300\ \mathrm{s}^{-1}$, and $\mu = 0.1$. Clearly, femtocell utilization is 0.1. Assume that femtocell is serving one UE. The probability that new service request arrives is $\frac{3\lambda_u}{3\lambda_u+\mu} = 0.09$. Considering utilization level, femtocells need to share transmission capacity about 1% of time. Therefore, we assumed that in underutilized network, assuming $C = 1$, will not change energy-delay tradeoffs. Finally, we leave analysys of different allocation schemes such as processor shared queue as a future work.

Now, assume that SBS is able to serve infinitely many UEs simultaneousy by sharing its transmission capacity. In this case, UE finds available SBS within its range immediately or in shorter time at the expense of long service periods. Optimizing number of UEs to be served simultenously with respect to traffic load can be investigated as a future work.

3.3 Relationship between Tolerable Delay and State Probability Distributions

Tolerable delay is not a parameter in the Markov model shown in Figure 3.1. Tolerable delay, w_t, is a deterministic value. Actual access delay however is random value between 0 and w_t with nonzero mass at edge points. In this part, we argue that probability of being idle, sleep or active is invariant with respect to tolerable delay. To show that we start with a Theorem. Then, we show simulation results that justify our proof.

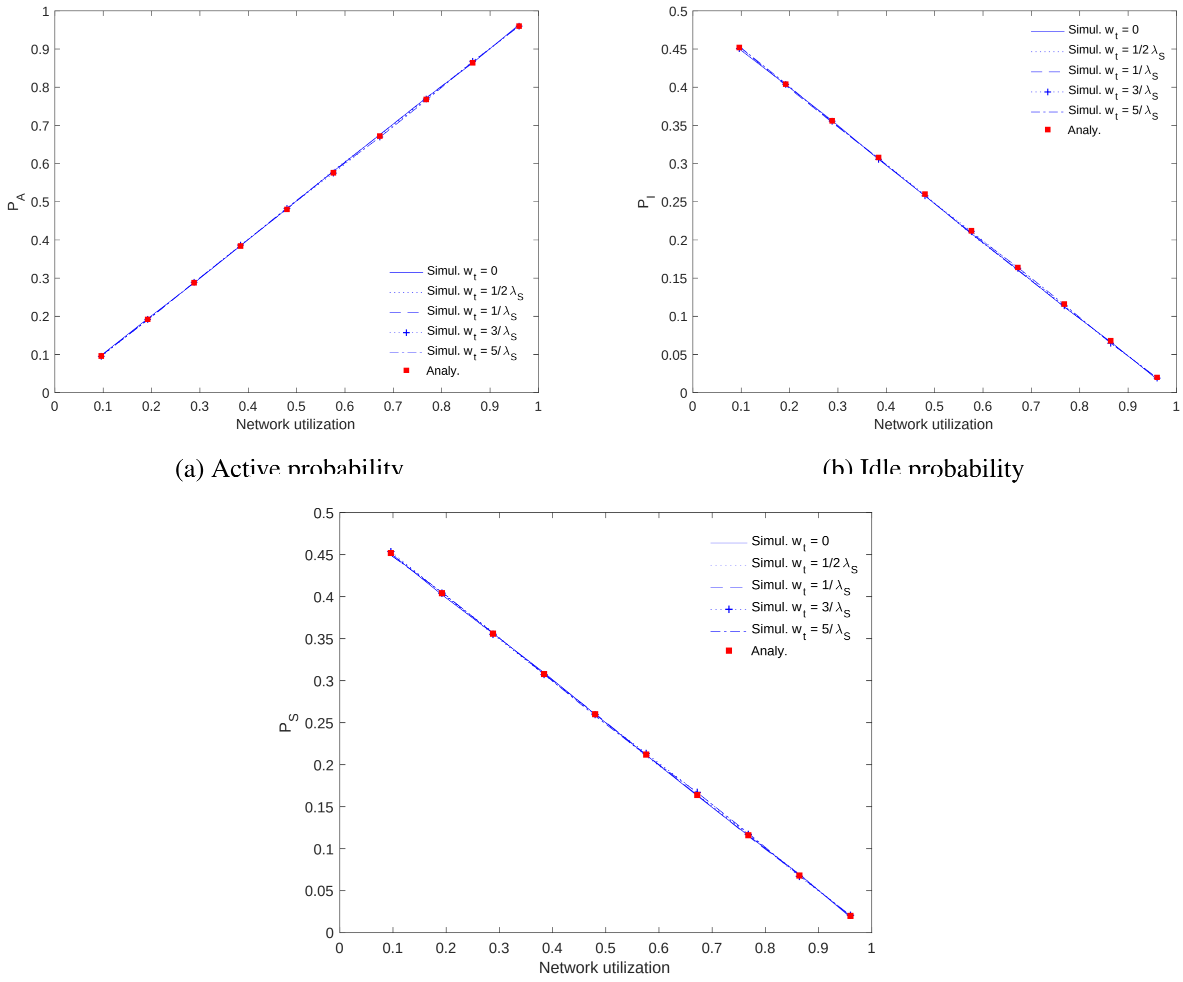
Theorem 1. *Let R be a Poisson process and τ be a random shift value with unknown distribution, with a range between 0 and W ($\tau \in [0, W]$). Then, the shifted process $R(\tau)$ still has a Poisson distribution.*

Proof. Without loss of generality, assume τ takes discrete and finite number of values, N, between 0 and W, where $\delta\tau = \frac{W}{N}$,$\tau_0 = 0$, $\tau_i = i\delta\tau$. Assume also that $\tau = \tau_1$ with probability p_1, and $\tau = \tau_2$ with probability p_2 and so on, and $\sum_{i=0}^{N} p_i = 1$. In this case, shifted process $R(\tau)$ is combination of $N + 1$ shifted processes. (i.e. $R(\tau) = R(\tau_0) \cup R(\tau_1) .. \cup R(\tau_N)$. Since the Poisson process is stationary, each of the shifted processes are Poisson, and overall process is combination of Poisson processes, so $R(\tau)$ is still Poisson. Regardless of the size of N, the process is still sum of shifted Poisson processes. □

Let t be the waiting time for a UE before receiving service. There are three possible cases for t depending on the access time of the data equipment: *i*) UE accesses to the cell immediately ($t = 0$), *ii*) UE accesses to a cell after tolerable delay time expires ($t = w_\mathrm{t}$), *iii*) UE has non-zero waiting time that is less than maximum tolerable delay ($0 < t < w_\mathrm{t}$). In the first case, $w_\mathrm{t} = 0$ means that it is not a parameter in the Markov chain since arrivals immediately receive service. Therefore, the model holds. In the second case, access time of UE is delayed by w_t time. In other words, case II is a shifted process of case I. Since Poisson process is stationary, the arrival distribution in case II is still Poisson. Finally, by Theorem 1, case III is also Poisson.

It is important to note that independent and random arrival locations will restore independence of arrival times, thereby omitting w_t in the Markov model. Delaying a connection changes distance distribution (not the service time distribution). If the user cannot connect to a small cell within w_t time units, it will still connect to a further small cell in the network. With the delayed connection, users will have the opportunity to connect to a closer small cell in the network, but service time in both cases will still be exponential with $1/\mu$. Therefore, network utilization or probability of cell being active will not change by the access time of the UE.

Figure 3.4 shows that state probabilities are invariant to waiting time. Threshold distance is chosen such that there are on average 5 cells around each UE. Mean sleep time, idle time, and service time are 10 seconds. Simulation results are repeated for various sleep rates and waiting times. We observe that w_t does not change the state probabilities.



(a) Active probability

(b) Idle probability

(c) Sleep probability

Figure 3.4: Independence of network utilization from waiting time of UE. Simulation parameters: $\rho_{\mathrm{u}} = 0.0012\,/m^2$, $\rho_{\mathrm{c}} = 0.0003/m^2$, $\mu = 0.1\,sec^{-1}$, $\lambda_{\mathrm{I}} = 0.1\,sec^{-1}$, $\lambda_{\mathrm{S}} = 0.1\,sec^{-1}$, $C = 1$.

3.4 Analysis of Transmit Distance Distributions

In this section, we analyze the distance distributions in rapid and delayed access schemes, and derive near-optimal threshold distance that minimizes average transmit distance of UE. We also give a simulation example to validate the analytical findings and the optimality condition for the threshold distance.

Obtaining the distance distributions is important because distance between a UE and a small cell determines the path loss, and thus small cell's transmit power. Using Friis's well-known transmission equation, transmit power of a small cell can be defined as

$$p_{\text{rx}} = \frac{H p_{\text{tx}}}{R^{\alpha}}, \tag{3.14}$$

where p_{tx}, p_{rx}, H, and α represent transmit power of an small cell, received power at UE, channel gain, and the path loss exponent, respectively. For constant p_{tx} and H, transmit power minimization problem solely depends on R^{α}. Therefore, minimizing R^{α} is the key point in minimizing the transmit power. In the remaining part of this section we analyze distribution of R^{α} in rapid and delayed access schemes.

3.4.1 Rapid Access

Analysis of distance distribution in rapid access scheme is straightforward. Using the access delay distribution in Eq. (3.13), probability that an small in radius R is immediately available is given by

$$P_{\text{in}}(R, 0) = 1 - e^{-\beta_0 \rho_c \pi R^2}. \tag{3.15}$$

Deriving the density function from Eq. (3.15),the mean of R^{α} is given by

$$\begin{aligned}\mathbb{E}\left[R^{\alpha}\right] &= 2\pi\beta_0\rho_{\mathrm{c}}\int_0^{\infty} r^{\alpha+1}e^{-\beta_0\rho_{\mathrm{c}}\pi r^2}\mathrm{d}r \\ &= (\beta_0\rho_{\mathrm{c}}\pi)^{-\frac{\alpha}{2}}\Gamma(1+\frac{\alpha}{2}).\end{aligned} \tag{3.16}$$

3.4.2 Delayed Access

Analysis of the distance distribution in delayed access scheme is more involved compared to rapid access. Let us consider two simple cases to understand how w_{t} affects the transmit distance. In the first case, w_{t} is very small or zero. Then, the UE will transmit immediately to the nearest available small cell, which is the same as in the rapid scheme. In the second case, w_{t} is sufficiently large so that the UE will be able to connect to a small cell within the range of R_{th}, if there is at least one, even if the small cell is initially unavailable. Regardless of how large w_{t} is, if the UE is able to access a small cell within R_{th}, transmit distance is R_{in}, and otherwise R_{out}. Clearly, we have $R_{\mathrm{in}} < R_{\mathrm{th}} < R_{\mathrm{out}}$ (see Figure 3.2). In remaining parts of this section, we derive the mean of R_{in}^{α} and $R_{\mathrm{out}}^{\alpha}$ conditioned on R_{th}.

For the sake of clarity, we start with a simple case to derive $\mathbb{E}[R_{\mathrm{in}}^{\alpha}|\,R_{\mathrm{th}}]$. Assume there is one small cell in the vicinity of UE. Then, distance distribution can be derived as,

$$\mathbb{P}(R_{\mathrm{in}} < r|\,R_{\mathrm{th}}, n = 1) = \frac{r^2}{R_{\mathrm{th}}^2}. \tag{3.17}$$

To facilitate the analysis and capture a convex function of R_{th}, (3.17) can be used to derive an upper bound on the expected value of R_{th}^{α} in delayed access scheme. Using (3.17), expected value of R_{in}^{α} is given by

$$\mathbb{E}\left[\left(R_{\mathrm{in}}^{\alpha}|R_{\mathrm{th}}\right)|n = 1\right] = \int_0^{R_{\mathrm{th}}} \frac{2r^{\alpha+1}}{R_{\mathrm{th}}^2}dr = \frac{2R_{\mathrm{th}}^{\alpha}}{\alpha+2}. \tag{3.18}$$

Given that DE accesses a small cell within its threshold distance, we further condition on the access time t (i.e. elapsed time until access occurs). Then, $\mathbb{E}\left[R_{\text{in}}^{\alpha}|R_{\text{th}}\right]$ can be expressed as,

$$\begin{aligned}\mathbb{E}\left[R_{\text{in}}^{\alpha}|R_{\text{th}}\right] &= \mathbb{E}\left[\left(R_{\text{in}}^{\alpha}|R_{\text{th}}\right)|t=0\right]q_0\\ &\quad+\mathbb{E}\left[\left(R_{\text{in}}^{\alpha}|R_{\text{th}}\right)|t>0\right](1-q_0).\end{aligned} \tag{3.19}$$

where q_0 is the conditional probability that waiting time is zero, hence $q_0=\frac{P_{\text{in}}(R_{\text{th}},0)}{P_{\text{in}}(R_{\text{th}},w_{\text{t}})}$. Conditioning on the number of available small cell, we can rewrite (3.19) as,

$$\begin{aligned}\mathbb{E}\left[R_{\text{in}}^{\alpha}|R_{\text{th}}\right] &= q_0\sum_{n=1}^{\infty}\mathbb{E}\left[\left(\left(R_{\text{in}}^{\alpha}|R_{\text{th}}\right)|t=0\right)|n\right]\Lambda(n|n\geq 1)\\ &\quad+\mathbb{E}\left[\left(R_{\text{in}}^{\alpha}|R_{\text{th}}\right)|t>0\right](1-q_0)\\ &\leq q_0\mathbb{E}\left[\left(R_{\text{in}}^{\alpha}|R_{\text{th}}\right)|n=1\right]\sum_{n=1}^{\infty}\Lambda(n|n\geq 1)\\ &\quad+\mathbb{E}\left[\left(R_{\text{in}}^{\alpha}|R_{\text{th}}\right)|n=1\right](1-q_0)\\ &=\frac{2R_{\text{th}}^{\alpha}}{\alpha+2}\end{aligned} \tag{3.20}$$

where $\Lambda(n|n\geq 1)=\frac{(\beta_0\rho_{\text{c}}\pi R_{\text{th}}^2)^n e^{-\beta_0\rho_{\text{c}}\pi R_{\text{th}}^2}}{n!(1-e^{-\beta_0\rho_{\text{c}}\pi R_{\text{th}}^2})}$. Inequality holds since $\mathbb{E}\left[\left(R_{\text{in}}^{\alpha}|R_{\text{th}}\right)|n=1\right]\geq\mathbb{E}\left[\left(R_{\text{in}}^{\alpha}|R_{\text{th}}\right)|n>1\right]$.

To find the mean of R_{out}^{α}, we consider a circle with radius $y>R_{\text{th}}$ in which an available small cell exists. Small cell exists in the area between two circles with radii R_{th}, and y. Then, the distance distribution of R_{out} is given by,

$$\mathbb{P}\left(R<y|\,R>R_{\text{th}}\right)=P_{\text{out}}=1-e^{-\beta_0\rho_{\text{c}}\pi(y^2-R_{\text{th}}^2)}. \tag{3.21}$$

Deriving density function from (3.21), the expected value of R_{out}^{α} can be found as

$$\mathbb{E}[R_{\text{out}}^{\alpha}|R_{\text{th}}]=\int_{R_{\text{th}}}^{\infty}y^{\alpha}d(P_{\text{out}})=\frac{e^{\beta_0 v}}{(\pi\beta_0\rho_{\text{c}})^{\frac{\alpha}{2}}}\Gamma(1+\frac{\alpha}{2},\beta_0 v), \tag{3.22}$$

where $v=\rho_{\text{c}}\pi R_{\text{th}}^2$, and $\Gamma(s,x)=\int_x^{\infty}t^{s-1}e^{-t}\text{d}t$.

3.4.3 Optimization of the Threshold Distance

Optimization of threshold distance is one of the key contributions in the paper. Expected value of R^{α} conditioned on R_{th} is given by

$$\mathbb{E}[R^{\alpha}|R_{\text{th}}] = \mathbb{E}[R_{in}^{\alpha}|R_{th}]P_{\text{in}} + \mathbb{E}[R_{\text{out}}^{\alpha}|R_{\text{th}}](1 - P_{\text{in}}). \tag{3.23}$$

Deriving the optimal R_{th} that minimizes (3.23) is cumbersome. However, it is possible to obtain near optimal R_{th} in closed form for $\alpha = 2$. Rewriting (3.23) for $\alpha = 2$, we have

$$\begin{aligned}\mathbb{E}[R^{2}|R_{\text{th}}] &= \mathbb{E}[R_{\text{in}}^{2}|R_{\text{th}}]P_{\text{in}}(R_{\text{th}}, w_{\text{t}}) \\ &\quad + \mathbb{E}[R_{\text{out}}^{2}|R_{\text{th}}]\big(1 - P_{\text{in}}(R_{\text{th}}, w_{\text{t}})\big) \\ &\leq \frac{R_{\text{th}}^{2}}{2}P_{\text{in}}(R_{\text{th}}, w_{\text{t}}) \\ &\quad + (R_{\text{th}}^{2} + \frac{1}{\beta_0 \pi \rho_{\text{c}}})\big(1 - P_{\text{in}}(R_{\text{th}}, w_{\text{t}})\big).\end{aligned} \tag{3.24}$$

$\mathbb{E}[R_{\text{out}}^{2}|R_{\text{th}}]$ can be easily derived from (3.22). The inequality is due to the bound in (3.20). After substituting $x \leftarrow R_{\text{th}}^{2}$, $\gamma \leftarrow \beta_{w_{\text{t}}}\pi\rho_{\text{c}}$, $\phi \leftarrow \frac{1}{\beta_0\pi\rho_{\text{c}}}$ in (3.24), we have

$$\mathbb{E}[x] = \frac{x}{2}(1 - e^{-x\gamma}) + (x + \phi)e^{-x\gamma}. \tag{3.25}$$

Taking the derivative of (3.25) with respect to x, substituting $e^{x\gamma} \approx 1 + x\gamma + \frac{(x\gamma)^2}{2}$, and equating to zero yields

$$R_{\text{th}}^{*} \approx \sqrt{\frac{2}{\beta_{w_{\text{t}}}\pi\rho_{\text{c}}}\sqrt{\frac{\beta_{w_{\text{t}}}}{\beta_0} - 1}}, \tag{3.26}$$

where R_{th}^{*} denotes Taylor approximation to an optimal threshold distance. Taking second derivative of (3.25) with respect to x, and substituting back x, γ, ϕ, one can show that second derivative is positive-definite. Therefore, global minimum exists for the upper bound we derived for (3.24). It is expected that R_{th}^* is sub-optimal due to the Taylor approximation, and the upper bound in (3.20).

Figure 3.5 shows the results for R_{th}^* . We observe that (3.26) is reasonably close to the optimum regardless of path loss exponent, and significantly improves the expected path loss. A gap between simulation and approximate analysis is due to the upper bound in Eq. (3.20).

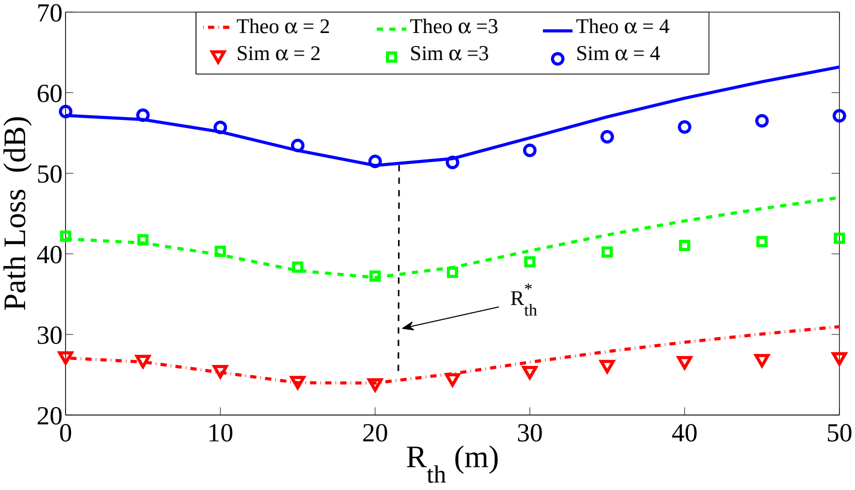


Figure 3.5: Comparison of the theoretical and simulated results for the expected path loss versus the threshold distance R_{th}. Dashed vertical line is Taylor approximation for R_{th}^*. Simulation parameters: $\rho_{\text{u}} = 0.003$, $\rho_{\text{c}} = 0.003$ /m^2, $\lambda_{\text{S}} = \frac{1}{\mu} = w_{\text{t}} = 0.1$ sec^{-1}, $\lambda_{\text{I}} = 0.33$ sec^{-1} $\lambda_{\text{u}} = 0.01$ sec^{-1}

3.5 Energy Efficiency Enhancements and Optimality Conditions

A UE reduces its transmit distance to a small cell by the delayed access scheme. However, it is still questionable whether or not UE can connect the nearest small cell in the delayed access scheme. In other words, it is necessary to discuss the bounds on transmit power savings in delayed access scheme because choosing the optimal threshold distance may not guarantee accessing the nearest small cell even if the tolerable delay is sufficiently large. Considering average transmit distance when all small cells are active as the optimum condition for UE small cell distance, we find how close to the optimum the average transmit distance in our delayed access scheme is.

For a given path loss exponent α, we define transmit power gain $G(\alpha)$, which is the ratio of the transmit power of UE with and without delay. Transmit power is minimum if $\beta_0 = 1$ in (3.12), i.e. all small cells are available for service. Transmit distance is maximum when w_{t} is zero and $\beta_0 < 1$, i.e. only a fraction of small cells is available. Comparing transmit distances for $\beta_0 = 1$, and $\beta_0 < 1$ for $w_{\mathrm{t}} = 0$ gives us an upper bound on $G(\alpha)$. Plugging (3.16) in (3.14) for both cases and taking their ratio yields

$$G(\alpha) \leq \frac{(\overline{p}_{\mathrm{tx}})_{\mathrm{rapid}}}{(\overline{p}_{\mathrm{tx}})_{\mathrm{min}}} = \left(\frac{1}{\beta_0}\right)^{\frac{\alpha}{2}}. \tag{3.27}$$

For the lower bound on $G(\alpha)$, we consider transmit power in rapid access and delayed access schemes. Plugging (3.16) and (3.23) in (3.14), and taking their ratio yields

$$G(\alpha) = \frac{(\overline{p}_{\mathrm{tx}})_{\mathrm{rapid}}}{(\overline{p}_{\mathrm{tx}})_{\mathrm{delayed}}} = \frac{(\mathbb{E}[R^\alpha])}{(\mathbb{E}[R^\alpha|R_{\mathrm{th}}])}. \tag{3.28}$$

Since $G(\alpha)$ increase with α. We consider free space as a special case (i.e., $\alpha = 2$). Plugging Eq. (3.26) in Eq. (3.28) yields

$$G(2) \geq \left[\beta_0 \left(e^{-2\hbar}\left[1 + (\frac{2}{\beta_{w_{\mathrm{t}}}} - 1)\hbar\right] + \hbar\right)\right]^{-1}, \tag{3.29}$$

where $\hbar = \sqrt{\frac{\beta_{w_{\mathrm{t}}}}{\beta_0} - 1}$. Inequality is due to Eq. (3.24). For $w_{\mathrm{t}} \to \infty$, Eq. (3.29) becomes

$$\frac{1}{\beta_0}\left(\underbrace{\left(e^{-2\sqrt{\beta_0^{-1}-1}} + 1\right)\left(\sqrt{\beta_0^{-1} - 1} + 1\right) - 1}_{\geq 1}\right)^{-1}, \tag{3.30}$$

hence $G(2) \leq \frac{1}{\beta_0}$. Eq. (3.30) implies that our delayed access scheme diverges from optimality as the proportion of available small cells in the network decreases. This is expected since the sleep times are exponential, and a UE has no information about an small cell's on/off schedule. Thus, delayed access scheme is not expected to yield optimal gains.

Result that we obtained in Eq. (3.30) also indicates how much the frequency reuse can be improved in the network with the delayed access since for $\alpha = 2$, R^2 is a unit measure for area. Considering this, our delayed access scheme decreases interference by decreasing average transmit power. We will discuss this in detail in Chapter 5.

3.6 Results and Discussions

In this section, we studied the performance of random on/off scheduling of small cells. We evaluated the effect of access delay on network power consumption and transmission power of UE. In these evaluations, we considered rapid access as the baseline, and compared its power savings with the delayed access. We developed a discrete-time simulator in MATLAB to validate our analytical model. We consider low network utilization (i.e., $\frac{\rho_{\mathrm{u}}\lambda_{\mathrm{u}}}{\rho_{\mathrm{c}}\mu} \leq 0.1$) with $\rho_{\mathrm{c}} = 0.005\,\mathrm{m}^{-2}$, $\rho_{\mathrm{u}} = 0.005\,\mathrm{m}^{-2}$, $\rho_{\mathrm{u}} = 0.005\,\mathrm{m}^{-2}$, $\lambda_{\mathrm{I}} = 0.01\,\mathrm{s}^{-1}$, $\lambda_{\mathrm{u}} = 0.01\,\mathrm{s}^{-1}$, $1/\mu = 10\,\mathrm{s}^{-1}$. Parameters are chosen to reflect high small cell density with respect to UE density. For example, $\rho_{\mathrm{u}} = 0.005$ indicates that mean distance to small cell is $\frac{1}{\sqrt{\rho_{\mathrm{u}}}} \approx 14.14\,m$, which can be interpreted as a dense femtocell deployment scenario. Arrival rate and service rate parameters are chosen so that network utilization is low. Unlike voice traffic parameters used in [81], we consider traffic pattern that has frequent inter-arrival and short session times. The rationale behind this choice is that background traffic and data traffic is more likely occur compared to voice traffic. Similar results can be obtained with different set of parameter values because comparison metrics in all scenarios, as we will explain, do not depend on parameter values. For the sake of generality and graph clarity, useful normalizations are also made.

We investigate how much small cell power consumption decreases as the tolerable delay increases while providing bounded transmit distance with high probability. To have a meaningful comparison, we consider the rapid access mode as a baseline where R_{th} is chosen so that $P_{\mathrm{in}} = 0.99$ and $w_{\mathrm{t}} = 0$. After that, for the same R_{th} we increase both the fraction of off small cells in the network and the tolerable delay subject to the constraint that transmit distance of UE is guaranteed, with high probability, to be within the same threshold distance. As a small cell, we consider femtocell. We used power consumption model for femtocell in Section 2.3. Power levels in sleep, idle, and active modes are 5.8 W, 9.6 W, and 10.4 W respectively.

Figure 3.6 shows that average power consumption of SBS decreases from 9.6 W to 6.3 W, corresponding to 86% of achievable savings by turning off RF unit, and 35% of overall power consumption compared to rapid access mode. If we decrease access probability P_{in}, required tolerable delay to satisfy the probability constraint decreases. Hence, lower power consumption level is achieved with less tolerable delay. We also observe in Figure 3.6 that tolerable delay for $P_{\text{in}} = 0.99$ has long tail. This is due to exponential distribution. By designing on/off schedule with deterministic rather than random sleep times, UE can access small cell with less tolerable delay, which is left as future work.

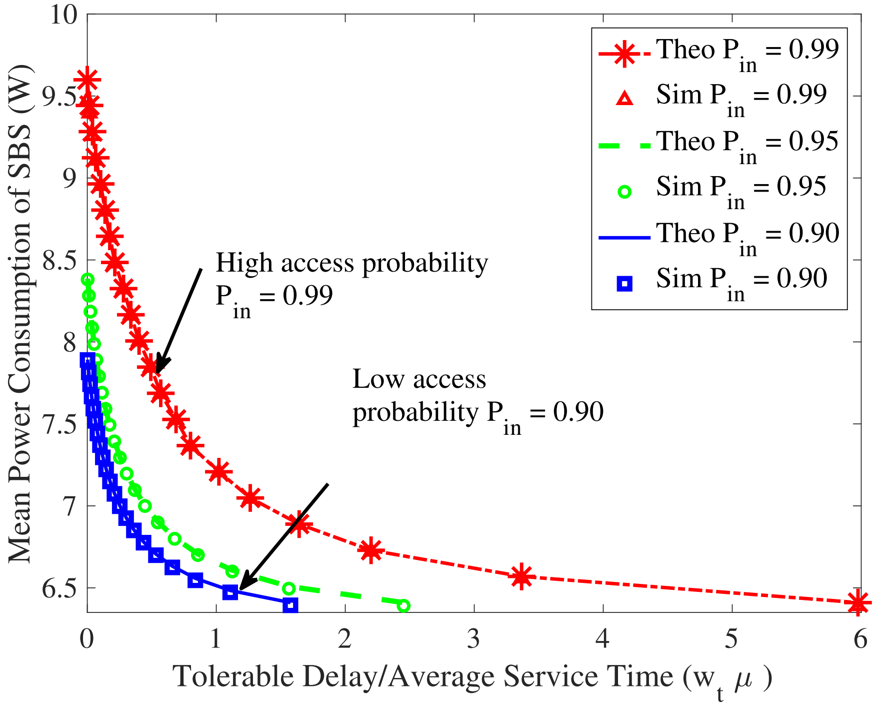


Figure 3.6: Variation of average power consumption of SBS at different access probabilities. Threshold distance is the same in all scenarios.

In Figure 3.7, for a given ratio of sleeping small cells and tolerable delay, we show how much the transmit power decreases by the delayed access scheme. We consider three scenarios. First scenario is *random on/off* mode, where small cells turn on and off in a short period. Second scenario is *permanent off* mode, where sleeping small cells are static and does not change during the simulation. Third scenario is *the permanent off mode with hexagonal small cell topology*. This scenario is the same as the second scenario except that the small cell topology is hexagonal. The reason to simulate hexagonal topology is to compare delayed access scheme with the case in which UE-to-cell distance is minimized. Fraction of sleeping small cells is 80% in all scenarios.

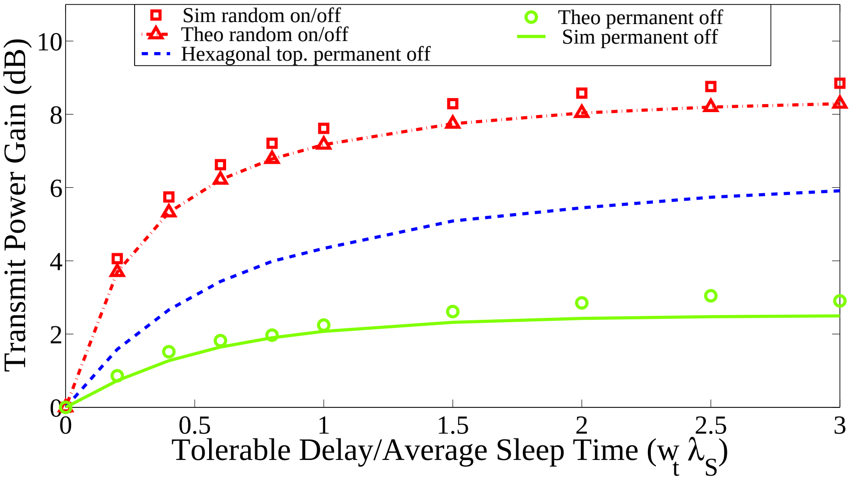


Figure 3.7: Transmit power gain with respect to tolerable delay with path loss exponent $\alpha = 3$. Small cells lie at the center of hexagonal tessellation. To keep number of SBSs in random and hexagonal topology the same, side length of each hexagon is chosen as $l_h = \sqrt{\frac{2}{3\sqrt{3}\rho_c}}$.

Results in Figure 3.7 show that the delayed access scheme decreases the transmit power by an order of magnitude. We observe that the transmit power gain in the first scenario is higher than the other scenarios. The hexagonal topology yields higher transmit power gains than random topology since average distance between UE and small cell is minimized. The optimal gain curve represents maximum achievable transmit power gains. In optimal case, all small cells are active hence average transmit distance is minimum. Delayed access scheme in all scenarios does not guarantee optimal transmit power gains because UE may access not *the closest* but *any* small cell within its threshold distance irrespective of the length of tolerable delay. To achieve optimal transmit power gains, an intelligent scheme is required in which an small cell shares its on/off schedule with the UEs in centralized or distributed way.

3.7 Conclusions and Future Work

Various cell access schemes and energy saving algorithms can be obtained by simply changing some of the model parameters, and decision criteria such as having deterministic sleep times instead of random sleep periods, or turning off a cell that is least likely to be busy instead of a randomly chosen one. Similarly, by changing the cell selection methods, different access schemes can be developed. In the baseline access scheme, access decision only depends on threshold distance, and waiting time. Taking into account the file size to be offloaded, or remaining waiting time and estimated service time of data equipment, new cell selection techniques can be devised.

In this paper, we propose delayed access mechanism to improve energy efficiency of small cells, and verify its performance via simulations. For the proposed access mechanism, optimal threshold distance minimizing average distance between a DE and small cell is derived. For the derived optimal threshold distance, further optimality conditions for transmit power and limitations of our delayed access scheme are discussed. Random and hexagonal topologies are used to demonstrate effectiveness of the proposed access scheme. Results show that the power consumption of small cells can be decreased by 35%, and the antenna transmit power of small cells can be decreased by several orders of magnitude by allowing initial access delays. Some of our future work include development on/off schemes with deterministic sleep times, and development of energy efficient collaboration protocols for small cells.

Chapter 4: Load Based Sleep Algorithms For Small Cell Networks

Massive densification of small cell networks (SCNs) is commonly seen as one of the major pillars of 5G wireless networks to cope with the ever-increasing mobile data traffic [82, 83]. For such dense deployments of SCNs, developing dynamic cell management and user-access mechanisms are crucial for saving energy at off-peak hours and for boosting the throughput of the network [84, 85]. Active cells not only consume energy, but also cause interference in the communication environment. Therefore, green and energy-efficient strategies that opportunistically place cells into sleep mode becomes important for irregular cell locations, especially with dynamically varying user distributions, spatial load, and traffic load.

In Chapter 3, we discussed simple solution approach to energy efficiency problem in detail. Instead of leaving cells off during off-peak hours, we proposed changing sleeping small cells dynamically and taking advantage of delay budget of UE. We showed that both average transmit distance, and average network power consumption can be decreased. In this chapter, we will argue that proper energy conserving schemes can be developed by estimating cell utilization, taking advantage of hardwave flexibility of small cells and delay tolerance of UE.

Siomina and Yuan [86] modeled cell load as coupled non-linear function of UE-cell distance, fading, interference, UE's service demand, and load of neighboring cells. The properties of load model are analytically derived, and numerical results are demonstrated for hexagonal topology. Fehske and Fettweis [87] found formulation for cell utilization that considers location of users and exponential service times. Similarly, Implicit formulation of the cell load is given, and energy saving scheme based on traffic load is designed in [88]. In all these approaches, UE-cell association is assumed to be static. While models are successful in accurately modeling utilization level of cell, they may not be the best option in designing low-complexity energy saving schemes for highly dense and complex random networks. As the network topology becomes large, implicit and detailed load formulations may bring computational overhead if they are used as design tool for on/off decision in energy saving algorithms.

It is expected that dense deployment and random distribution of small cells will lead to coverage overlapping, where UE may be in communication range of multiple small cells to offload data. Therefore, it is possible to take advantage of overlapping areas and shut off the cells with the lowest expected traffic utilization, especially at off-peak hours. Such overlapping also occurs in macrocell tier. Marsan et al. [14] proposed energy saving operation of two neighboring towers that belong to different operators and have overlapping area. However, their analytical approach is limited to a pair towers.To design energy saving algorithms specifically for small cells, new load models are necessary that consider not only the overlapping coverage, but also UE's traffic allocation to nearby cells due to changing UE-cell associations in sleeping modes.

Flexibility of small cells ease the realization of energy-conversing schemes. There is a rich literature about energy-efficiency of small cells. Sumudu et al. [85] proposed game theoric approach to improve energy efficiency of ultra dense small cell networks. Li et al. [89] gave energy efficiency analysis of small cells. Their model considers both small cells and macrocells. Relationship between energy-efficiency and number of transmit antennas of base stations is studied. Soh et al. [90] studied energy efficiency of small cell network with random and strategich sleep modes that is based on traffic load. Merwaday and Guvenc [91] studied energy efficiency and spectral efficiency of small cell and marcocell network which operates enhances interference coordiation, and range expansion. However, energy savings can be further improved by dynamic switching based on short-term service demand, and integrating energy-efficient sleep mode techniques with flexible access strategies for UEs. For example, in [92], the geographical area is divided into multiple grids. In each grid area, a maximum number of SBSs is selected at times of peak traffic to satisfactorily serve all users. In idle periods, a subset of these selected SBSs is kept active and remaining SBSs are turned off. This strategy yields up to 53% energy savings in dense areas, and 23% in sparse areas. The difference is due to fact that number of cells that can be turned off in dense areas are more compared with sparse areas.

Delay tolerance of UE gives additional options for turning SBSs on and off. For example, depending on the delay budget a central controller can turn on a nearby sleeping cell. While delay tolerant networks (DTN) have been studied extensively in the literature in [93, 94, 95], it has not been explored well in the context of energy-efficient SCNs, where placing certain small cells into sleep mode can save energy at the cost of latency for certain users. In [96], traffic arriving at single cell from multiple UEs are modeled as M/G/1 vacation queue with close-down and setup times. Queue is analyzed rigorously, and optimal sleeping policy is designed. On-off scheme based on accumulated tasks is proposed, and its delay-energy efficiency aspects are analyzed. Instead of single cell, this chapter aims to explore UE's delay tolerance in small cell network operating in energy efficient manner.

This chapter is a rigorously extended version of [97, 98]. It studies energy-efficient on/off scheduling (OOS) strategies for SBSs in next-generation 5G networks. Considering a user-centric approach, a similar access mechanism discussed previously (see Figure 3.3) we propose a novel *load* based OOS framework with a promise of more energy-efficient SCNs. Our main contributions are:

- We propose a simple effective traffic load metric for dense SCNs. The traffic load of the overall network is represented by a random *load* variable. We investigate the distribution of this load variable, and derived analytical expressions of respective probability distribution function (PDF) and cumulative distribution function (CDF), which are verified through extensive simulations.

- Towards achieving energy-efficient SCNs, we propose two load based (LB) OOS algorithms, where certain fraction of SBSs with relatively lower load values are put into less energy consuming (i.e., sleeping) states for a random duration of time. In particular, we introduce *centralized* LB (CLB) and *distributed* LB (DLB) as two novel on/off scheduling algorithms. Although CLB needs the knowledge of instantaneous load values of all SBSs, DLB, instead, relies on the CDF of the load requiring much fewer load samples. The numerical results verify that CLB and its computational-efficient alternative DLB have very close performance.

– We also consider two benchmark OOS techniques, which are *random on/off* (ROO) and *wake-up control* (WUC), in which central controller is assumed to have complete information about on/off schedules and delay budget of UEs, and makes on/off decisions accordingly. While ROO is a simple baseline algorithm [42], WUC is a more complex sophisticated algorithm requiring full-control of the macro base station (MBS) dynamically. The numerical results verify that CLB and DLB are superior to ROO, and have similar performance as WUC. Furthermore, as the overall SCN traffic increases, WUC turns out to be less energy-efficient than either CLB or DLB.

The rest of the chapter is organized as follows. Section 4.1 introduces the system model for SCNs with dynamic on/off operation of SBSs. Section 4.2 analytically derives the traffic load distribution for a given UE using a Gamma distribution approximation. Section 4.3 proposes the centralized and distributed strategies to conduct on/off operation of SBSs. Section 4.5 presents numerical results, and Section 4.6 concludes the chapter.

4.1 System Model

In this section, we first review the network model, then describe the novel load based model for the network traffic, and finally describe the power consumption model of the SBSs.

4.1.1 SCN Model

We consider a very similar model to the one used in Chapter 3, with some differences. We consider a densely packed SCN where low-power SBSs are operated to deliver mobile data to UEs of interest. UEs generate traffic at random time intervals, and request to offload a file where the file size and the service request intervals have exponential distribution with rates λ_{F} and λ_{U}. This chapter intoduces file size parameter which was not used in previous chapter. Having said that, any file size distribution can be assumed without making any change in the model. We will discuss this in Section 4.2.4

Considering that each UE is not involved in transmission all the time (due to exponentially distributed service request times), the energy efficiency of the overall network can be improved by putting some of the SBSs into less energy-consuming (i.e., sleeping) states. Leaving details of sleeping states and the associated OOS strategies to Section 4.3, each SBS in sleeping states is assigned with a random sleep time T_s, which follows exponential distribution with rate λ_S. A slightly different delayed access strategy under consideration is given in Figure 4.1, which assumes that any UE has tolerable delay of at most w_t seconds (i.e., *waiting time*). If a UE with active service request finds at least one available (i.e., *idle*) SBS within the threshold distance R_th and the waiting time w_t, it connects to the best (i.e., *nearest*) of these SBSs to offload its desired traffic. Otherwise, it connects to MBS, and the current service request is assumed to be *blocked* at the SCN level. In terms of interaction between UEs and SBSs, we assume that UEs do not know the location of sleeping SBSs. But rather, UEs have the perfect knowledge of distance to each non-sleeping SBSs, which can be estimated by monitoring and processing the downlink reference signals from these SBSs. The association between UEs and SBSs is set up such that each SBS serves a single UE at a time, and each UE does not change its SBS till the current service request is completely fulfilled. Association multiple UEs with single small cell while considering the amount of data to be transferred may be advantageous if the size of files tobe offloaded are small. Meantime, if UEs that request service are located too closely, association to different SBSs lead to interfere unless a central controller has careful frequency allocation strategy in SCN. However, if UEs are not too close, association to different SBSs with delayed acccess have positive effect spectral efficiency. At this point, more sophisticated UE-cell associations can be made by considering location of UEs, frequency assigment strategy, file size to be transferred at each session, and delay budget of UE. We leave this problem as a future work, and we assume that in underutilized network small cells mostly engages with one user at a time. In addition, UEs use all available bandwidth once connected to an SBS, and quickly finish their service resulting in short service times. We finally note that SCN handles only the data traffic, and the voice traffic is handled efficiently by MBSs in macrocell tier.

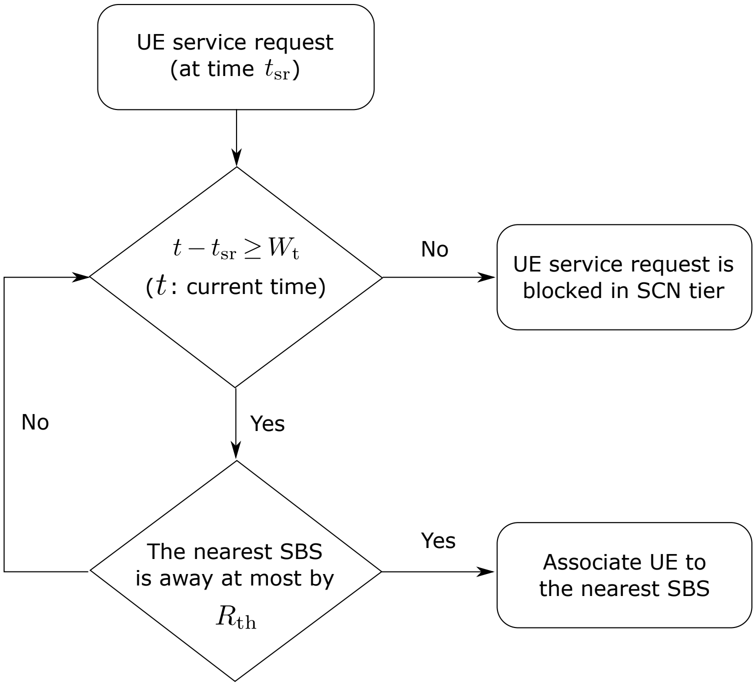

UE service request
(at time t_{sr})
$t - t_{\mathrm{sr}} \geq W_{\mathrm{t}}$
(t: current time)
No
UE service request is blocked in SCN tier
No
Yes
The nearest SBS is away at most by R_{th}
Yes
Associate UE to the nearest SBS

Figure 4.1: Delayed access strategy of UEs in SCN. If there is no cell available in range, UE keeps waiting. If a cell becomes available within R_{th} transmission begins. Otherwise, service request is blocked.

4.1.2 SBS/UE Densities and Traffic Load

In this section, a new load metric is introduced. Key design consideration of the load metric is that it allocates UE's traffic load not only to the nearest cells but also a possible set of SBSs that UE may connect within its communication range R_{th} as SBS are randomly distributed and their availability changes due to being fully active or being in sleep mode. We define n_{c} and n_{u} as range-dependent SBS and UE densities, respectively, which refer to average number of SBSs and UEs within a circular area of radius R_{th}. Since location distribution for SBSs and UEs both follow HPPP, the respective Poisson distribution with the range-dependent SBS and UE densities are defined with the mean values $\nu_{\mathrm{c}} = \rho_{\mathrm{c}}\pi R_{\mathrm{th}}^2$ and $\nu_{\mathrm{u}} = \rho_{\mathrm{u}}\pi R_{\mathrm{th}}^2$, respectively. The probability that n_{c} SBSs and n_{u} UEs are present in the circular area of radius R_{th} are therefore given as $p_{\mathrm{c}}(i) = P\{n_{\mathrm{c}} = i\} = \frac{\nu_{\mathrm{c}}^i e^{-\nu_{\mathrm{c}}}}{i!}$ and $p_{\mathrm{u}}(j) = P\{n_{\mathrm{u}} = j\} = \frac{\nu_{\mathrm{u}}^j e^{-\nu_{\mathrm{u}}}}{j!}$, respectively.

We define the *load factor* for the jth UE as follows:

$$w_j = \begin{cases} \frac{1}{n(j)} & \text{if } n(j) > 0, \\ 0 & \text{if } n(j) = 0, \end{cases} \tag{4.1}$$

where $n(j)$ is the number of SBSs that the jth UE can receive service (i.e., away by at most R_{th}). Accordingly, *load value* L_i for the ith SBS is defined to be the sum of load factors associated with each UE off the ith SBS by at most a distance of R_{th}, and is given as follows:

$$L_i = \sum_{j=1}^{\infty} w_j \mathbb{1}(i, j), \tag{4.2}$$

where $\mathbb{1}(i, j)$ is the indicator function which is 1 if ith SBS and jth UE are within R_{th} distance, and zero otherwise.

As an example, we consider a representative network given in Figure 4.2. Defining $\mathcal{S}_i$ as the indices of SBSs that ith UE can receive service, we have $\mathcal{S}_1 = \{1\}$, $\mathcal{S}_2 = \{1,2\}$, $\mathcal{S}_3 = \{1,2,3\}$, and $\mathcal{S}_4 = \{2,3\}$. Using (4.1), load factors of UEs are computed as $w_1 = 1$, $w_2 = \frac{1}{2}$, $w_3 = \frac{1}{3}$, $w_4 = \frac{1}{2}$. The respective load values of SBSs are then given using (4.2) by $L_1 = 1 + \frac{1}{2} + \frac{1}{3} = \frac{11}{6}$, $L_2 = \frac{1}{2} + \frac{1}{3} + \frac{1}{2} = \frac{4}{3}$, and $L_3 = \frac{1}{3} + \frac{1}{2} = \frac{5}{6}$. It is important to note that actual average traffic load to the cell i is $\lambda_{\text{U}} \lambda_{\text{F}} L_i$. Since network is homogeneous (i.e. arrival rate and file size distributions are same for all UEs), simplified load metric in (4.2) is sufficient.

We note that instantaneous load value of any SBS possibly varies with the blocked calls, UE-SBS association policies, traffic patterns, and transmission rates in a particular UE-SBS topology. Therefore, the load value of a SBS may not represent exact load distribution perfectly, but is successful enough in giving a good measure of how much traffic any specific SBS handles. Last but not least, the load computation in given example in Figure 4.2 is completely obtained in distributed manner. Analysis of load distribution allows design of robust distributed sleep mode algorithms, which will be discussed thoroughly in following sections.

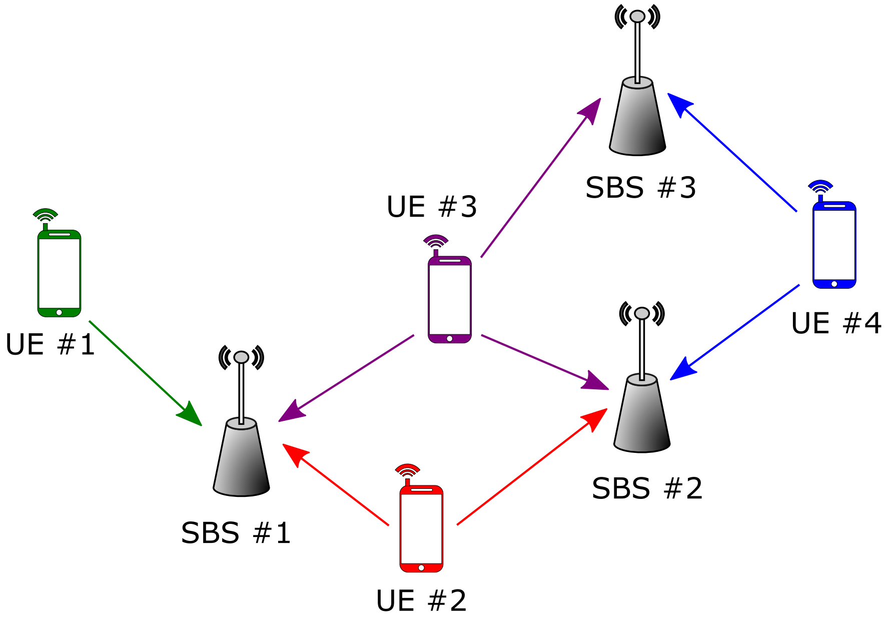


Figure 4.2: A representative network of 3 SBSs and 4 UEs. The arrows indicate the SBSs that each UEs can receive service.

4.1.3 Power Consumption Model for SBSs

We now briefly present the power consumption model for an arbitrary SBS, which is an important measure while evaluating the energy efficiency of the overall network. Note that since the transmission range of UEs is very limited, and respective delayed access strategy described in Fig 4.1 is the same for all OOS strategies, average transmit power of UE is expected to be invariant in all schemes. We therefore do not include the power consumption of UEs in this study, and take into account power consumption of SBSs only.

Considering a standard BS architecture, we assume that the hardware is composed of three blocks: microprocessor (i.e., to manage radio protocols, backhaul connection, etc.), field-programmable gate array (FPGA) (i.e., to process necessary baseband algorithms), radio frequency (RF) front-end (e.g., power amplifiers, transmitter elements, etc.) [42, 99, 100]. In order to obtain power saving (i.e., *sleeping*) states, OOS strategies consider to turn off a fraction of SBSs not actively engaged in transmission. This can be done by turning off some or all of the hardware blocks, where it takes more time to boot up as more hardware blocks are turned off (i.e., deeper the state is).

In Table 4.1, we list the SBS states considered in this work together with respective boot-up

Table 4.1: SBS States, Boot-Up Times, and Power Consumption Levels

SBS State	Boot-up time (s)	Power Consumption (%)
Active	0	100
Idle	0	50
Standby	0.5	50
Sleep	10	15
Off	30	0

times and normalized power consumption levels, which are available in the literature, [42, 99, 100] The description and assumptions for the SBS states are as follows.

- *Active*: The SBS is actively engaging in transmission with full power.
- *Idle*: The SBS is ready to transmit immediately, but not transmitting currently. Hence, RF front-end is not running, and the power consumption is therefore 50% of active state.
- *Standby*: In this light sleep state, the heater for oscillator is turned off intentionally, and RF front-end is not running at all.
- *Sleep*: The SBS is in a deep sleep with only necessary hardware parts (power supply, central processor unit (CPU), etc.) are up.
- *Off*: The SBS is completely offline.

Note that, the sleeping state should be put into either sleep or off states to achieve significant power savings, where the respective minimum boot-up time is 10 seconds. Since any sleeping SBS should be available right after its random sleep time T_s expires, it is not possible to put any SBS into either sleep or off states if $T_\mathrm{s} < 10$ seconds. We assume that such SBSs are put into standby state, as shown in Table 4.2, to capture the effect of turning off procedure, and meet the requirement to wake up immediately after T_s seconds. In addition, the power consumption during boot-up period is equal to that of the standby state since that particular SBS does not actively communicate with users.

Although deeper sleeping states provide more power savings, respective longer boot-up times result in UE service requests being blocked more in SCN tier. To effectively handle this fundamental trade-off between energy consumption and boot-up time, the optimal sleep state should be selected based on UE's delay tolerance, transmit range, and cell density. The decision of optimal state and sleep duration is beyond the scope of this study. Instead, we prefer a simple rule which puts each SBS into the deepest state as much as possible for maximum power savings and is given below. Sleep times up to 30 seconds are stand-by or sleep which are determined by hardware limitations. However, if the sleep time is greater than 30 seconds, then, SBS can be either in sleep or off mode. Decision between sleep or off mode is made by minimum power consumption rule by taking into consideration of both the power consumption during boot-up, $p_{\text{boot-up}}$, and sleep mode p_{sleep}.

Table 4.2: Sleep State Choice Based on Sleep Time (T_{s})

Sleep State	Sleep Time (T_{s}) (s)
Stand-by	$T_{\text{s}} \leq 10$
Sleep	$10 < T_{\text{s}} \leq 30$
Sleep	$T_{\text{s}} > 30, 10\, p_{\text{boot-up}} + (T_{\text{s}} - 10)\ p_{\text{sleep}} < 30\, p_{\text{boot-up}}$
Off	$T_{\text{s}} > 30, 10\, p_{\text{boot-up}} + (T_{\text{s}} - 10)\ p_{\text{sleep}} > 30 p_{\text{boot-up}}$

4.2 Analysis of Traffic Load Distribution

In this section, we analyze the distribution of the load variable as a successful measure of the actual traffic loads of SBSs. There are several studies in the literature where fitting distributions are used instead of deriving exact distributions, especially for Poisson Voronoi cell topologies [101, 102, 103]. Following a similar approach, we analyze distribution of the load variable L by considering the Gamma distribution, which is verified to have satisfactory fitting performance.

The PDF of the gamma distribution can be expressed in terms of shape parameter k_{l} and scale parameter θ_{l} as follows:

$$f(x; k_{\text{l}}, \theta_{\text{l}}) = \frac{\theta_{\text{l}}^{k_{\text{l}}} x^{k_{\text{l}}-1} e^{-\theta_{\text{l}} x}}{\Gamma(k_{\text{l}})}, \tag{4.3}$$

where $\Gamma(\cdot)$ is the gamma function [104]. Our goal is, therefore, to determine suitable expressions of the gamma parameters k_{l} and θ_{l} in terms of SCN parameters ρ_{u}, ρ_{c}, and R_{th}. When the load variable L is assumed to be gamma-distributed with parameters k_{l} and θ_{l}, the first and second moments are given as

$$\mathbb{E}[L] = \frac{k_{\mathrm{l}}}{\theta_{\mathrm{l}}}, \qquad \mathbb{E}[L^2] = \frac{k_{\mathrm{l}}\,(1 + k_{\mathrm{l}})}{\theta_{\mathrm{l}}^2}, \tag{4.4}$$

and the parameters to be determined can be expressed as

$$k_{\mathrm{l}} = \theta_{\mathrm{l}}\mathbb{E}[L], \tag{4.5}$$

$$\theta_{\mathrm{l}} = \frac{\mathbb{E}[L]}{\mathbb{E}[L^2] - \mathbb{E}[L]^2}. \tag{4.6}$$

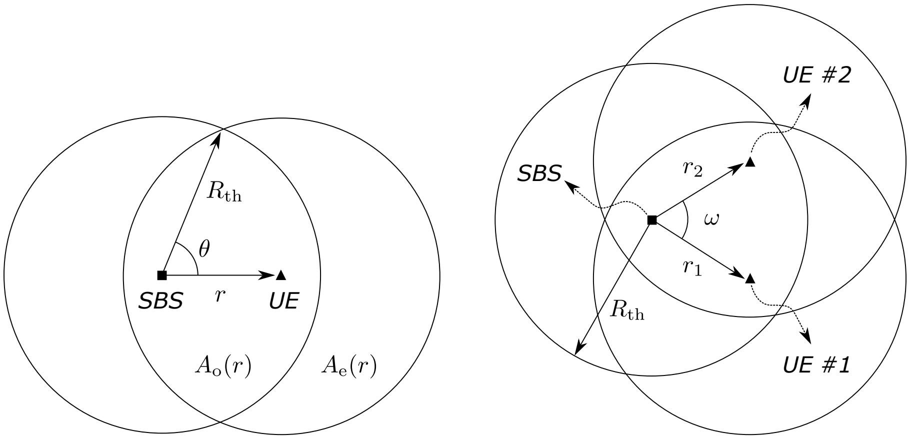
As a result, the first and the second moments of L completely specifies the desired fitting distribution, and the rest of our analysis is therefore devoted to finding these moments.

4.2.1 First Moment of Load Variable

The first moment of the load variable L for arbitrary SBS in the network is derived by focusing on a representative sub-network shown in Figure 4.3(a). In this framework, the target SBS (for which the load will be computed) is assumed to be located at the origin together with n_{c} additional SBSs and n_{u} UEs, which are distributed randomly over a circular area of radius R_{th}.

The first moment of the load L can be expressed as a conditional sum over all possible number of SBSs and UEs as follows:

$$\mathbb{E}\,[L] = \sum_{i=0}^{\infty}\sum_{j=0}^{\infty} \mathbb{E}\,[L \mid n_{\mathrm{c}} = i, n_{\mathrm{u}} = j]\; p_{\mathrm{c}}(i)\, p_{\mathrm{u}}(j), \tag{4.7}$$



(a) Single UE

(b) Two UEs

Figure 4.3: A representative SCN involving a single SBS at the origin, and arbitrary UEs off by at most $R_{\rm th}$.

and using the load definition of (4.2) in (4.7) yields

$$\mathbb{E}\,[L] = \sum_{i=0}^{\infty}\sum_{j=0}^{\infty}\mathbb{E}\left[\sum_{k=1}^{j} w_k | n_{\rm c} = i\right] p_{\rm c}(i)\, p_{\rm u}(j), \tag{4.8}$$

$$= \sum_{i=0}^{\infty}\sum_{j=0}^{\infty}\sum_{k=1}^{j}\mathbb{E}[w_k | n_{\rm c} = i]\, p_{\rm c}(i)\, p_{\rm u}(j). \tag{4.9}$$

We observe that the individual load factors in (4.9) (i.e., w_k's) are not necessarily the same since the number of SBSs which are away from each UE by at most $R_{\rm th}$ may not be the same. The expected values of the load factors are, however, the same (i.e., $\mathbb{E}[w_k|n_{\rm c}=i]=\mathbb{E}[w|n_{\rm c}=i]$ for $\forall\, k$) since SBSs follow homogeneous Poisson Point process . We may therefore rearrange (4.9) to obtain

$$\mathbb{E}\,[L] = \sum_{i=0}^{\infty}\sum_{j=0}^{\infty} j\,\mathbb{E}[w | n_{\rm c} = i]\, p_{\rm c}(i)\, p_{\rm u}(j), \tag{4.10}$$

$$= \sum_{i=0}^{\infty}\mathbb{E}[w | n_{\rm c} = i]\, p_{\rm c}(i) \sum_{j=1}^{\infty} j\, p_{\rm u}(j). \tag{4.11}$$

Realizing that the last summation in (4.11) is the definition of the expected value for the number of users (i.e., n_u), which is Poisson distributed with rate ν_u, we obtain

$$\mathbb{E}\left[L\right] = \nu_\mathrm{u} \sum_{i=0}^{\infty} \mathbb{E}[w|n_\mathrm{c} = i]\, p_\mathrm{c}(i), \tag{4.12}$$

which reduces to the problem of finding *average traffic load* contributed by a *single UE*.

In order to compute the average load factor conditioned on the number of cell (i.e., $\mathbb{E}[w|n_\mathrm{c} = i]$), we choose an arbitrary UE that is off the origin (i.e., the target SBS of interest) by a distance r with $r \leq R_\mathrm{th}$, as shown in Figure 4.3(a). Because any user can only receive service from the cells separated by at most a distance of R_th, the cells that contribute into the load factor are those lying in the overlapping area $A_\mathrm{o}(r)$ and the user exclusion area $A_\mathrm{e}(r)$, as shown in Figure 4.3(a). These areas can be expressed parametrically as follows

$$A_\mathrm{o}(r) = 2r^2 - \theta + \frac{1}{2}\sin(2\theta), \tag{4.13}$$

$$A_\mathrm{e}(r) = \pi R_\mathrm{th}^2 - A_\mathrm{o}(r), \tag{4.14}$$

where $\theta = \cos^{-1}\left(\frac{r}{2R_\mathrm{th}}\right)$ is also depicted in Figure 4.3(a).

The conditional load factor involved in (4.12) could be expressed as follows

$$\mathbb{E}[w|n_\mathrm{c} = i] = \int_0^{R_\mathrm{th}} \mathbb{E}[w|r, n_\mathrm{c} = i] f_r(r)\mathrm{d}r, \tag{4.15}$$

where $f_r(r) = 2r/R_\mathrm{th}$. The average load factor in (4.15), which is conditioned on the distance r and the number of cells i (i.e., located within a circle of radius R_th around the origin), can be expressed as a sum in the form of a binomial expansion as follows

$$\mathbb{E}[w|r, n_\mathrm{c} = i] = \sum_{k=0}^{i} \binom{i}{k} \mathbb{E}[w|r, n_{A_\mathrm{o}}(r) = k] p_{A_\mathrm{o}}(r)^k (1 - p_{A_\mathrm{o}}(r))^{i-k}, \tag{4.16}$$

where $n_{A_\mathrm{o}}(r)$ stands for the number of cells in the overlapping area $A_\mathrm{o}(r)$, and $p_{A_\mathrm{o}}(r)$ is the probability of an SBS being in $A_\mathrm{o}(r)$. Since SBSs follow Poisson distribution, we have $p_{A_\mathrm{o}}(r) = A_\mathrm{o}(r)/\pi R_\mathrm{th}^2$.

In addition, each term in the summation of (4.16) considers a case in which k SBSs exist in the overlapping area $A_{\mathrm{o}}(r)$ out of a total of i SBSs off the origin by at most the distance R_{th}.

While computing the average load expression at the right side of (4.16) by employing the definition given in (4.1), one should take into account k SBSs from the overlapping area $A_{\mathrm{o}}(r)$, v SBSs from the user exclusion area $A_{\mathrm{e}}(r)$, and the single cell located at the origin as follows

$$\begin{aligned}\mathbb{E}[w|r, n_{A_{\mathrm{o}}(r)} = k] &= \sum_{v=0}^{\infty} \frac{1}{k+v+1} P\left\{n_{A_{\mathrm{e}}}(r) = v\right\} \\ &= \sum_{v=0}^{\infty} \frac{[\nu_e(r)]^v \, e^{-\nu_e(r)}}{(k+v+1)\, v!},\end{aligned} \tag{4.17}$$

where $n_{A_{\mathrm{e}}}(r)$ is the random variable representing the number of SBSs in the user exclusion area $A_{\mathrm{e}}(r)$, which follows the Poisson distribution with rate $\nu_{\mathrm{e}}(r) = \rho_{\mathrm{c}} A_{\mathrm{e}}(r) = \nu_{\mathrm{c}} - \rho_{\mathrm{c}} A_{\mathrm{o}}(r)$. Finally, employing (4.15)-(4.17) and $f_r(r) = 2r/R_{\mathrm{th}}$ in (4.12), the first moment of L is obtained as follows:

$$\mathbb{E}[L] = \frac{2\nu_{\mathrm{u}}}{R_{\mathrm{th}}} \sum_{v=0}^{\infty} \sum_{i=0}^{\infty} \sum_{k=0}^{i} \frac{\nu_{\mathrm{c}}^i e^{-\nu_{\mathrm{c}}}}{(k+v+1) i! v!} \binom{i}{k} \int_0^{R_{\mathrm{th}}} [\nu_e(r)]^v \, e^{-\nu_e(r)} p_{A_{\mathrm{o}}}(r)^k (1-p_{A_{\mathrm{o}}}(r))^{i-k} r \mathrm{d}r, \tag{4.18}$$

which is a function of the UE density ν_{u}, the SBS density ν_{c}, and the threshold distance R_{th}.

4.2.2 Second Moment of Load Variable

Following the same approach of (4.7), the second moment of L can be written as

$$\begin{aligned}\mathbb{E}\left[L^2\right] &= \sum_{i=0}^{\infty} \sum_{j=0}^{\infty} \mathbb{E}\left[L^2 | n_{\mathrm{c}} = i, n_{\mathrm{u}} = j\right] p_{\mathrm{c}}(i)\, p_{\mathrm{u}}(j), \\ &= \sum_{i=0}^{\infty} \sum_{j=0}^{\infty} \mathbb{E}\left[\left(\sum_{k=1}^{j} w_k\right)^2 \middle| n_{\mathrm{c}} = i\right] p_{\mathrm{c}}(i)\, p_{\mathrm{u}}(j),\end{aligned} \tag{4.19}$$

which can be written after some manipulation as follows

$$\begin{aligned}\mathbb{E}\left[L^2\right] &= \sum_{i=0}^{\infty}\sum_{j=0}^{\infty}\left(\sum_{k=1}^{j}\mathbb{E}\left[w_k^2|n_{\mathrm{c}}=i\right] + \sum_{k=1}^{j}\sum_{\substack{l=1\\ l\neq k}}^{j}\mathbb{E}[w_k w_l|n_{\mathrm{c}}=i]\right)p_{\mathrm{c}}(i)\,p_{\mathrm{u}}(j)\\ &= \sum_{i=0}^{\infty}\sum_{j=0}^{\infty}\left(j\,\mathbb{E}\left[w^2|n_{\mathrm{c}}=i\right] + j(j-1)\,\mathbb{E}\left[w_k w_l|n_{\mathrm{c}}=i\right]\right)p_{\mathrm{c}}(i)p_{\mathrm{u}}(j) \qquad (4.20)\end{aligned}$$

for any k, l with $k \neq l$. Following the discussion in obtaining (4.12) from (4.11), and employing first and second-order statistics of the Poisson distribution, we have

$$\sum_{j=1}^{\infty} j(j-1)\,p_{\mathrm{u}}(j) = \mathbb{E}\left[n_{\mathrm{u}}^2\right] - \mathbb{E}\left[n_{\mathrm{u}}\right] = {\nu_{\mathrm{u}}}^2, \qquad (4.21)$$

and (4.20) accordingly becomes

$$\mathbb{E}\left[L^2\right] = \underbrace{\nu_{\mathrm{u}}\sum_{i=0}^{\infty}\mathbb{E}\left[w^2|n_{\mathrm{c}}=i\right]p_{\mathrm{c}}(i)}_{E_1} + \underbrace{\nu_{\mathrm{u}}^2\sum_{i=0}^{\infty}\mathbb{E}\left[w_k w_l|n_{\mathrm{c}}=i\right]p_{\mathrm{c}}(i)}_{E_2}. \qquad (4.22)$$

The expectation $E[w^2|n_{\mathrm{c}} = i]$ in (4.22) can be computed following the steps of (4.15)-(4.17) together with the modified version of (4.17) given as

$$\mathbb{E}[w^2|r, n_{A_{\mathrm{o}}}(r) = k] = \sum_{v=0}^{\infty}\frac{[\nu_{\mathrm{e}}(r)]^v\, e^{-\nu_{\mathrm{e}}(r)}}{(k+v+1)^2\, v!}, \qquad (4.23)$$

and the first expression at the right hand side of (4.22) becomes

$$E_1 = \frac{2\nu_{\mathrm{u}}}{R_{\mathrm{th}}}\sum_{v=0}^{\infty}\sum_{i=0}^{\infty}\sum_{k=0}^{i}\frac{\nu_{\mathrm{c}}^i e^{-\nu_{\mathrm{c}}}}{(k+v+1)^2 i!v!}\binom{i}{k}\int_0^{R_{\mathrm{th}}}[\nu_e(r)]^v\, e^{-\nu_e(r)}p_{A_{\mathrm{o}}}(r)^k(1-p_{A_{\mathrm{o}}}(r))^{i-k}r\mathrm{d}r. \qquad (4.24)$$

However, computation of the second expectation in (4.22) is cumbersome due to the correlation between the individual load factors w_k and w_l.

Because the expectation $\mathbb{E}[w_k w_l|n_{\mathrm{c}} = i]$ requires a second-degree analysis, we modify Figure 4.3(a) by adding a second user, and obtain Figure 4.3(b). This new coordinate system has a

SBS located at the origin, as before, and two UEs off this cell by random distances r_1 and r_2, both of which have the common distribution with $f_r(r) = 2r/R_{\text{th}}$. We may have various orientations for relative positions of two UEs in Figure 4.3(b), and therefore introduce a new variable ω which describes the difference of user angles with respect to the origin.

Note that ω is actually the difference of two uniform random variables distributed between 0 and 2π. The distribution of ω is therefore given as [105]

$$g(\omega) = \begin{cases} \dfrac{\omega}{2\pi(2\pi+1)} & \text{if } \omega \in [-2\pi, 0], \\ \dfrac{1-\omega}{4\pi^2} & \text{if } \omega \in [0, +2\pi]. \end{cases} \tag{4.25}$$

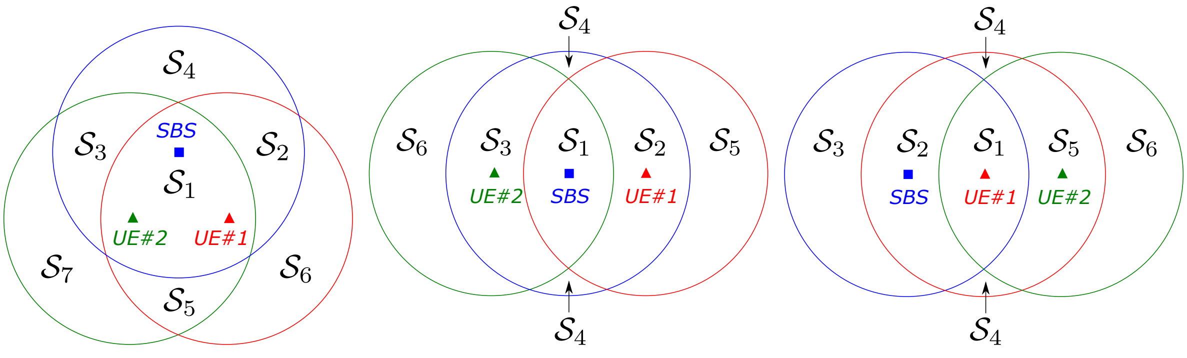
The second-order expectation of interest could be accordingly written as

$$\mathbb{E}[w_k w_l | n_{\text{c}} = i] = \int_0^{R_{\text{th}}} \int_0^{R_{\text{th}}} \int_{-2\pi}^{2\pi} E\left[w_k w_l \,|\, \boldsymbol{r}, \omega, n_{\text{c}} = i\right] f_r(r_1)\, f_r(r_2)\, g(\omega)\, \mathrm{d}\omega\, \mathrm{d}r_1\, \mathrm{d}r_2, \tag{4.26}$$

which is counterpart of (4.15) in the first moment computation, and where $\boldsymbol{r} = [r_1 \;\; r_2]$.

In order to compute the expectation at the right side of (4.26), we need to consider various geometric orientations of two UEs around the origin, as in Figure 4.4. Among them, Case-I has a circular triangular overlapping area whereas Case-II and Case-III specify non-triangular overlapping areas. While the condition for the existence of a circular triangle area and respective area formulations are given in [106], the non-triangular areas should be computed by employing (4.13).

In order to express the term $E[w_k w_l | \boldsymbol{r}, \omega, n_{\text{c}} = i]$ in the form of multinomial expansion, we need to take into account the number of constituent areas (i.e., N) forming the circular area of radius R_{th} around the origin (i.e., where the SBS is located). Note that the expectation in (4.26) assumes $i+1$ SBSs in this circular region. Indeed, N is a function of the angle ω given in Figure 4.3(b), and all 3 cases sketched in Figure 4.4 occurs for a certain set of ω values [106]. Based on these 3 orientations in Figure 4.4, Case-I and Case-II have $N = 4$ constituent areas while Case-III has



(a) Case-I (b) Case-II (c) Case-III

Figure 4.4: Relative orientations of two UEs around a single SBS.

$N = 3$. As a counterpart of (4.16), the desired expansion could therefore be given as

$$E[w_k w_l | \boldsymbol{r}, \omega, n_c = i] = \sum_{m_1=0}^{i} \cdots \sum_{m_{N-1}=0}^{i - \sum_{v=0}^{N-2} m_v} E[w_k w_l | \boldsymbol{r}, \omega, \boldsymbol{n}(\boldsymbol{r}, w) = \boldsymbol{m}] f(\boldsymbol{m}; \boldsymbol{p}(\boldsymbol{r}, w)), \tag{4.27}$$

where $\boldsymbol{n}(\boldsymbol{r}, w)$ is the vector of the number of SBSs in each of the constituent areas, $\boldsymbol{p}(\boldsymbol{r}, w)$ is the vector of multinomial probabilities associated with each of these areas, and $\boldsymbol{m}$ is the vector of summation indices. Each term of the summation in (4.27) corresponds to a unique distribution of the total of i SBSs over the constituent areas. Specifically, the number of SBSs in the constituent area $A_v(\boldsymbol{r}, w)$ is $n_v(\boldsymbol{r}, w) = m_v$ for $v = 1, 2, \ldots, N$ with $\sum_{v=1}^{N} m_v = i$.

The probability mass function (PMF) in (4.27) is given as

$$f\left(\boldsymbol{m}; \boldsymbol{p}(\boldsymbol{r}, w)\right) = i! \prod_{v=1}^{N} (m_v!)^{-1} \prod_{v=1}^{N} p_{A_v}(\boldsymbol{r}, w)^{m_v}, \tag{4.28}$$

where $p_{A_v}(\boldsymbol{r}, w)$ is the individual probability entry of $\boldsymbol{p}(\boldsymbol{r}, w)$ associated with the constituent area $A_v(\boldsymbol{r}, w)$, and is therefore given to be $p_{A_v}(\boldsymbol{r}, w) = A_v(\boldsymbol{r}, w)/\pi R_{\text{th}}^2$ owing to the uniform distribution of SBSs in space. Note that m_v SBSs in $A_v(\boldsymbol{r}, w)$ can be placed in $m_v!$ different ways, and this makes $\prod_{v=1}^{N} m_v!$ considering all constituent areas. Since the total of i SBSs can be ordered in $i!$ different ways, $i! \prod_{v=1}^{N} (m_v!)^{-1}$ in (4.28) takes into account all possible relative SBS placements.

Following the philosophy behind (4.17), and employing the PMF in (4.28), the expectation in the summation of (4.27) can be computed as follows

$$E[w_k w_l | \boldsymbol{r}, \omega, \boldsymbol{n}(\boldsymbol{r}, w)] = \sum_{v_1=0}^{\infty}\sum_{v_2=0}^{\infty}\sum_{v_c=0}^{\infty} P\left\{n_{\mathrm{e,c}}(\boldsymbol{r}, w) = v_{\mathrm{c}}\right\} \prod_{s=1}^{2} \frac{P\left\{n_{\mathrm{e},s}(\boldsymbol{r}, w) = v_s\right\}}{n_{\mathrm{o},s}(\boldsymbol{r}, w) + v_s + v_{\mathrm{c}} + 1}, \tag{4.29}$$

$$= \sum_{v_1=0}^{\infty}\sum_{v_2=0}^{\infty}\sum_{v_c=0}^{\infty} \frac{\left[\nu_{\mathrm{e,c}}(r)\right]^{v_{\mathrm{c}}} e^{-\nu_{\mathrm{e,c}}(r)}}{v_{\mathrm{c}}!} \prod_{s=1}^{2} \frac{\left[\nu_{\mathrm{e},s}(r)\right]^{v_s} e^{-\nu_{\mathrm{e},s}(r)}}{v_{\mathrm{s}}!\left(n_{\mathrm{o},s}(\boldsymbol{r}, w) + v_s + v_{\mathrm{c}} + 1\right)}, \tag{4.30}$$

where $n_{\mathrm{e,c}}(\boldsymbol{r}, w)$ and $n_{\mathrm{e},s}(\boldsymbol{r}, w)$ are the number of SBSs in the common exclusion area $A_{\mathrm{e,c}}(\boldsymbol{r}, w)$ and distinct exclusion area $A_{\mathrm{e},v}(\boldsymbol{r}, w)$ for the sth UE, respectively, which follow the Poisson distribution with rates $\nu_{\mathrm{e,c}}(r) = \rho_{\mathrm{c}} A_{\mathrm{e,c}}(r)$ and $\nu_{\mathrm{e},s}(r) = \rho_{\mathrm{c}} A_{\mathrm{e},s}(r)$, respectively, with $s = 1, 2$. We show all the exclusion and overlapping areas in Table 4.3 for the orientations considered in Figure 4.4.

Table 4.3: Overlapping and Exclusion Areas

	Case-I	Case-II	Case-III
$A_{\mathrm{e,c}}(\boldsymbol{r}, w)$	$\mathcal{S}_5$		$\mathcal{S}_5$
$A_{\mathrm{o},1}(\boldsymbol{r}, w)$	$\mathcal{S}_1 \bigcup \mathcal{S}_2$	$\mathcal{S}_1 \bigcup \mathcal{S}_2$	$\mathcal{S}_1 \bigcup \mathcal{S}_2$
$A_{\mathrm{e},1}(\boldsymbol{r}, w)$	$\mathcal{S}_6$	$\mathcal{S}_5$	$\mathcal{S}_4$
$A_{\mathrm{o},2}(\boldsymbol{r}, w)$	$\mathcal{S}_1 \bigcup \mathcal{S}_3$	$\mathcal{S}_1 \bigcup \mathcal{S}_3$	$\mathcal{S}_1$
$A_{\mathrm{e},2}(\boldsymbol{r}, w)$	$\mathcal{S}_7$	$\mathcal{S}_6$	$\mathcal{S}_6$

Note that $n_{\mathrm{o},s}(\boldsymbol{r}, w)$ in (4.29) is a given (i.e., deterministic) value representing the number of SBSs in the overlapping area $A_{\mathrm{o},s}(\boldsymbol{r}, w)$, with $s = 1, 2$. More specifically, $n_{\mathrm{o},s}(\boldsymbol{r}, w)$ is the sum of the entries of $\boldsymbol{n}(\boldsymbol{r}, w)$ associated with the constituent areas forming $A_{\mathrm{o},s}(\boldsymbol{r}, w)$, which are explicitly given in Table 4.3 for $s = 1, 2$. As an example, we have $n_{\mathrm{o},1}(\boldsymbol{r}, w) = n_1(\boldsymbol{r}, w) + n_2(\boldsymbol{r}, w)$ and $n_{\mathrm{o},2}(\boldsymbol{r}, w) = n_1(\boldsymbol{r}, w) + n_3(\boldsymbol{r}, w)$ for Case-I, where $n_i(\boldsymbol{r}, w)$ is the number of SBSs in the area $\mathcal{S}_i$ for $i = 1, 2, 3$.

As a particular case, since $A_{\mathrm{e,c}}(\boldsymbol{r}, w)$ does not exist for Case-II, (4.30) simplifies to

$$E[w_k w_l | \boldsymbol{r}, \omega, \boldsymbol{n}(\boldsymbol{r}, w)] = \sum_{v_1=0}^{\infty}\sum_{v_2=0}^{\infty} \prod_{s=1}^{2} \frac{\left[\nu_{\mathrm{e},s}(r)\right]^{v_{\mathrm{s}}} e^{-\nu_{\mathrm{e},s}(r)}}{v_{\mathrm{s}}!\left(n_{\mathrm{o},s}(\boldsymbol{r}, w) + v_s + 1\right)}. \tag{4.31}$$

Combining (4.26)-(4.30), we finally obtain E_2 appearing in (4.22) as follows

$$E_2 = \frac{4\nu_{\mathrm{u}}^2}{R_{\mathrm{th}}^2} \sum_{i=0}^{\infty} \sum_{\nu_1=0}^{\infty} \sum_{\nu_2=0}^{\infty} \sum_{\nu_{\mathrm{c}}=0}^{\infty} \nu_{\mathrm{c}}^{i} e^{-\nu_{\mathrm{c}}} \int_{0}^{R_{\mathrm{th}}} \int_{0}^{R_{\mathrm{th}}} \int_{-2\pi}^{2\pi} \sum_{m_1=0}^{i} \dots \sum_{m_{N-1}=0}^{i-\sum_{v=0}^{N-2} m_v} \frac{\left[\nu_{\mathrm{e,c}}(r)\right]^{\nu_{\mathrm{c}}} e^{-\nu_{\mathrm{e,c}}(r)}}{\nu_{\mathrm{c}}!\, m_1! \dots m_N!} \prod_{v=1}^{N} p_{A_v}(\boldsymbol{r}, w)^{m_v}$$
$$\times \prod_{s=1}^{2} \frac{\left[\nu_{\mathrm{e},s}(r)\right]^{\nu_{\mathrm{s}}} e^{-\nu_{\mathrm{e},s}(r)}}{\nu_{\mathrm{s}}!\, \left(n_{\mathrm{o},s}(\boldsymbol{r}, w) + \nu_s + \nu_{\mathrm{c}} + 1\right)} g(\omega)\, r_1 r_2\, \mathrm{d}\omega\, \mathrm{d}r_1 \mathrm{d}r_2, \qquad (4.32)$$

which is also a function of densities ν_{u} and ν_{c}, and the distance R_{th}.

As a result, the respective parameters k_1 and θ_1 of the fitting gamma distribution can be computed using the first order moment $\mathbb{E}[L]$ given in (4.18), and the second order moment $\mathbb{E}[L^2]$ given in (4.22) (i.e., the sum of (4.24) and (4.32)), based on the relations given in (4.5) and (4.6). The CDF of load distribution can therefore be written as

$$F_L(x) = P\{L < x\} = e^{-\nu_{\mathrm{u}}} + (1 - e^{-\nu_{\mathrm{u}}}) \int_{0^+}^{x} \frac{\theta_1^{k_1}}{\Gamma(k_1)} y^{k_1 - 1} e^{-\theta_1 y} \mathrm{d}y, \qquad (4.33)$$

where first term represents the void probability, $P\{L = 0\}$ (i.e., no user is around the SBS). Using (4.33), the respective PDF of load distribution can be written as

$$f_L(x) = \begin{cases} e^{-\nu_{\mathrm{u}}} & \text{if } x = 0, \\ (1 - e^{-\nu_{\mathrm{u}}}) \dfrac{\theta_1^{k_1}}{\Gamma(k_1)} x^{k_1 - 1} e^{-\theta_1 x} & \text{if } x > 0. \end{cases} \qquad (4.34)$$

4.2.3 Effect of Localization Error

In practice, the UE-cell distance may not be measured accurately. In this section, we discuss the scenarios in which load distribution under erroneous measurements may still hold, or needs further analysis. We argue that as long as UE's spatial distribution remains HPPP, the load distribution holds. For example, assume the widely used Gaussian error, e_{L} with zero mean and standard deviation σ, ($e_{\mathrm{L}} \sim \mathcal{N}(0, \sigma)$), and update (x, y) locations as $(x + e_{\mathrm{L}}, y + e_{\mathrm{L}})$. A 2-dimensional Poisson Process can be considered combination of two 1-dimensional Poisson processes corresponding to

x- and x-coordinates, which are independent. Without loss of generality, assume for now location error is only on x-axis and constant l_e. Let p_{l_e} is the probability that error to occur. According to our error model, $p_{l_e} = \frac{1}{\sqrt{2\pi\sigma^2.}} e^{-\frac{x^2}{2\sigma^2}}$. Now, random x coordinates forms of two subsets of UEs with densities $p_{l_e}\rho_{\mathrm{u}}$, and $(1 - p_{l_e})\rho_{\mathrm{u}}$ corresponding to erroneous x coordinates, and perfectly measured x-coordinates respectively. Process of erroneous x coordinates is in fact a Poisson processes shifted by l_e, which is still Poisson due to stationarity of Poisson processes. It is clear that due to superposition property, the combined processes still form Poisson Process. The same reasoning applies regardless of the size of error. Finally, we note that in case the localization errors are correlated or dependent, the load distribution needs further analysis due to the fact that stationarity property of Poisson process no longer holds.

4.2.4 Modeling Traffic Load Under Diverse Traffic Patterns

Macrocells were initially planned to handle voice traffic, and motivation behind deployment of small cells is to meet increasing data traffic demand. In our model, we considered homogeneous traffic pattern. However, the simple load metric in the study may be still useful for modeling non-homogeneous traffic patterns. We discuss two cases: *i*) Each UE generate different traffic patterns but aggregate traffic load from UEs are same, *ii*) There are different type of UEs with high and low traffic demand. For the first case, we argue that our analysis still holds. On the other hand, second case requires further analysis.

We just give some insight here, and made appropriate changes to clarify the model. In case each UE generates two types of traffic, say, UE uploads short and long size files with means $\bar{F}$ $\bar{F}_2$ at λ_{U_1}, λ_{U_2} rates. From Erlang's perspective traffic is product of arrival rate and holding time. The holding time (i.e. service time) increases in proportion to file size. So, we can re-define the load factors of UE with $n(j)$ neighboring cells as $\frac{\lambda_{\mathrm{u}_1}\bar{F}_1}{n(j)}$, $\frac{\lambda_{\mathrm{u}_2}\bar{F}_2}{n(j)}$. Then, the aggregate traffic load factor of UE having $n(j)$ neighbors will be $\frac{\lambda_{\mathrm{u}_1}\bar{F}_1 + \lambda_{\mathrm{u}_2}\bar{F}_2}{n(j)}$. Since $\lambda_{\mathrm{u}_1}\bar{F}_1 + \lambda_{\mathrm{u}_2}\bar{F}_2$ is constant, it factors out from Eq. (4.2). Then, first and second moments of weighted traffic loads are $(\lambda_{\mathrm{u}_1}\bar{F}_1 + \lambda_{\mathrm{u}_2}\bar{F}_2)\mathbb{E}[L]$, and $(\lambda_{\mathrm{u}_1}\bar{F}_1 + \lambda_{\mathrm{u}_2}\bar{F}_2)^2\mathbb{E}[L^2]$. It is clear that with such approach, load distribution can still be obtained.

In case, there are different types of UEs such as intense users having high service demand with density ρ_{u_1}, and low profile UEs with density ρ_{u_2}, load metric still can be used. Let aggressive users have service request with rate λ_{u_1} with mean file size $\bar{F}_h$, and let low profile users have service requests λ_{u_2} with mean file size with mean file size $\bar{F}_l$. For each traffic types, fitting distribution parameters can be found using the similar method by applying weights $\rho_{u_1}\lambda_{\mathrm{u}_1}\bar{F}_h$, $\rho_{\mathrm{u}_2}\lambda_{\mathrm{u}_2}\bar{F}_l$. Then, tje sum of the two load distributions may be analyzed. Two distributions may be correlated due to common cell topology; therefore, calculation of the sum of the load distributions arising from each type of UE may be challenging problem.

4.3 Load Based On/Off Scheduling

In this section, we study on off scheduling (OOS) strategies with a goal of having more energy-efficient SCNs. In this respect, we first consider a random OOS algorithm (i.e., ROO) to set up a simple benchmark to evaluate performance of smarter OOS strategies. We then propose two novel load based OOS algorithms, which are called centralized locad based (CLB) and distributed load based (DLB), and establish a good compromise between energy-efficiency and network throughput. Finally, we also consider a more sophisticated OOS strategy, which is called wake-up control scheme (WUC), where the central controller has the full capability to wake up any sleeping SBSs.

We assume that the percentage of sleeping SBSs are fixed in all the OOS algorithms under consideration for the sake of a fair comparison. As a result, for each sleeping SBS to wake up, the OOS algorithms choose the *best idle* SBS to turn off. Depending on the specific OOS algorithm, the set of sleeping SBSs may dynamically change as the turn-off and turn-on events occur repeatedly.

We also assume that any UE can get service from the available SBSs, which are either *currently idle* or *become idle* within the waiting time period, as discussed in Section 4.1.1. In particular, ROO, CLB, and DLB strategies assume no capability at the central controller to wake up a sleeping SBS during its random sleep time. The WUC strategy, however, assumes that the central controller can give order to wake up a sleeping SBS to make it available within the waiting time (i.e., where it would otherwise not become available).

4.3.1 Random On/Off Scheduling

In this strategy, a central controller (e.g., macrocell) turns off randomly selected idle cell, and assigns a random sleep time for each SBS having been turned off. Each sleeping SBS wakes up automatically after its sleep time expires, and the central controller decides which SBS to turn off in return. The overall procedure is given in Algorithm 1.

Algorithm 1 Random On/Off Scheduling (ROO)

```
Input: The sleep time of $i$th SBS has expired
$\text{SBS}_{\text{nextToSleep}} \leftarrow \text{ROO}(i, \mathcal{S}_{\text{all}})$        ▷ $\mathcal{S}_{\text{all}}$ is the set of all SBSs
turn off $\text{SBS}_{\text{nextToSleep}}$
procedure ROO($i$,$\mathcal{S}_{\text{all}}$)        ▷ ROO algorithm
    $\mathcal{S}_{\text{idle}} \leftarrow \text{find}_{1 \le \ell \le |\mathcal{S}_{\text{all}}|}(\text{ state}(\mathcal{S}_{\text{all}}(\ell)) == \text{idle })$
    $j \leftarrow \text{rand}(1, |\mathcal{S}_{\text{idle}}|)$
    return $\mathcal{S}_{\text{idle}}(j)$
end procedure
```

4.3.2 Centralized Load Based (CLB) On/Off Scheduling

The centralized locad based (CLB) can be considered to be the load based alternative of random on/off (ROO), which operates in a centralized fashion as described in Algorithm 2. In CLB, the central controller turns off the SBS with the minimum instantaneous load value computed using Eq. (4.2) as a response to each SBS that has just waken up.

Note that the algorithm needs the load values of idle SBSs only to make on/off decision. Meantime, UE shares its instantaneous load factor with the idle and active cells because of two reasons: *i*) each active cell may return to idle status after completion of the transmission and may be available within w_t time, *ii*) density of non-sleeping cells (i.e., idle and active) do not change and therefore, distribution of load can be obtained, which allows implementation of on/off decision in distributed manner. Computing load distribution of only idle cells is very challenging since the set of idle cells dynamically change.

Algorithm 2 Centralized Load Based On/Off Scheduling (CLB)

```
**Input**: The sleep time of $i$th SBS has expired
$\text{SBS}_{\text{nextToSleep}} \leftarrow \text{CLB}(i, \mathcal{S}_{\text{all}})$        ▹ $\mathcal{S}_{\text{all}}$ is the set of all SBSs
turn off $\text{SBS}_{\text{nextToSleep}}$
**procedure** $\text{CLB}(i,\mathcal{S}_{\text{all}})$        ▹ CLB algorithm
    $\mathcal{S}_{\text{idle}} \leftarrow \text{find}_{1\le\ell\le|\mathcal{S}_{\text{all}}|}(\,\text{state}(\mathcal{S}_{\text{all}}(\ell)) == \text{idle}\,)$
    compute $L_\ell$ by (4.2) for $\ell = 1, \ldots, |\mathcal{S}_{\text{idle}}|$
    $j \leftarrow \text{argmin}_{1\le\ell\le|\mathcal{S}_{\text{idle}}|}\, L_\ell$
    **return** $\mathcal{S}_{\text{idle}}(j)$
**end procedure**
```

4.3.3 Distributed Load Based On/Off Scheduling

The DLB algorithm is a distributed version of the centralized CLB algorithm, where the overall operation does not need a central controller. In the DLB approach, whenever a sleeping SBS is about to wake up (i.e., after expiration of its random sleep time), that specific SBS is designated to be the decision-maker to decide the next SBS to be turned off. The decision-maker SBS first determines all its idle first-hop neighbours (i.e., within a distance of at most R_{th}) as the candidate SBSs to be turned off. The instantaneous load values of the candidate SBSs are then collected (e.g., via BS-BS communication using X2 backhaul link [107]), and the idle SBS with the minimum instantaneous load is chosen by the decision-maker SBS as the one to turn off next.

An important feature of DLB is the mechanism specifying when to stop searching candidates in a *wider* neighborhood. To this end, the algorithm checks the following relation

$$1 - (1 - P\{L < L_{\text{min}}\})^{|\text{S}(k+1)|} < \kappa \tag{4.35}$$

where L_{min} is the minimum instantaneous load associated among the cells traversed up to k hops, and κ is a threshold probability. Given the cardinality of $|S(k+1)|$ idle cells at next hop, $k+1$, (4.35) checks the probability of finding a cell with a lower load than that of L_{min}. Note that (4.35) can be computed readily using the analytical load CDF in (4.33). If inequality of (4.35) is correct, then the algorithm stops searching for a better candidate SBS, and decides to turn off the current candidate. Otherwise, the algorithm widens its search to second-hop neighbours (i.e., those in $2R_{\text{th}}$

distance). Likewise, algorithm continues to widen its search till it becomes less likely to find an SBS with a lower load than that of the existing candidate (i.e., for which (4.35) turns out to be true). The complete procedure is given in Algorithm 3.

Algorithm 3 Distributed Load Based On/Off Scheduling (DLB)

```
**Input**: The sleep time of $i$th SBS has expired
SBS$_{\text{nextToSleep}} \leftarrow$ DLB$(i, \kappa)$
turn off SBS$_{\text{nextToSleep}}$
**procedure** DLB$(i,\kappa)$                                              ▷ DLB algorithm
    $L_{\min} \leftarrow \infty$, $k \leftarrow 1$
    **while** $1 - (1 - F_L\,(L_{\min})\})^{|\mathcal{S}(k+1)|} > \kappa$ **do**
        $\mathcal{S} \leftarrow \text{find}_{1\le\ell\le|\mathcal{S}_{\text{all}}|}(\,\text{distance}(\mathcal{S}_{\text{all}}(\ell), \mathcal{S}_{\text{all}}(i)) \le kR_{\text{th}}\,)$      ▷ $\mathcal{S}_{\text{all}}$ is the set of all SBSs
        $\mathcal{S}_{\text{idle}} \leftarrow \text{find}_{1\le\ell\le|\mathcal{S}|}(\,\text{state}(\mathcal{S}(\ell)) == \text{idle}\,)$
        compute $L_\ell$ by for $\ell = 1, \ldots, |\mathcal{S}_{\text{idle}}|$
        $j \leftarrow \text{argmin}_{1\le\ell\le|\mathcal{S}_{\text{idle}}|}\, L_\ell$
        $L_{\min} = L_j$
        $k \leftarrow k + 1$
    **end while**
    **return** $\mathcal{S}_{\text{idle}}(j)$
**end procedure**
```

The distribution of load changes as the SBSs are turned on and off. Consider two networks where fixed proportion of SBSs are switched off randomly, and by load based on algorithms. Let L_R be random load variable of SBS operated by random on/off algorithm, and let L_{LD} random load variable of SBS operated by load based on/off. It is clear that

$$\begin{aligned} P\{L_{LD} < l\} < P\{L_R < l\} &\iff 1 - P\{L_{LD} < l\} > 1 - P\{L_R < l\} \\ &\iff (1 - P\{L_{LD} < l\})^{|S|} > (1 - P\{L_R < l\})^{|S|} \\ &\iff 1 - (1 - P\{L_{LD} < l\})^{|S|} < 1 - (1 - P\{L_R < l\})^{|S|} \end{aligned} \tag{4.36}$$

Therefore, stopping condition in (4.35) is still valid stopping condition for the networks operated by load based algorithms. Tighter bound on (4.35) can be found by utilizing order statistics of load distribution, which is left to a future work.

4.3.4 Wake-up Control Based On/Off Scheduling (WUC)

We finally consider a more complex approach, which is called wake-up control (WUC) and given in Algorithm 4. This algorithm is, indeed, very similar to the CLB algorithm, except that the central controller now has the full control to wake up any sleeping SBS, even before the respective sleep time expires. This way, any of the UE service requests, which could not otherwise be met by available idle SBSs, might be handled by incorporating the sleeping SBSs. To do so, the candidate sleeping SBSs should be within the communication range of the UE, and be able to wake up within the tolerable delay of that UE holding the current request. More specifically, the boot-up time of the candidate sleeping SBSs (given in Table 4.1) should end within the tolerable delay. Note that once the central controller places a wake-up order for the nearest candidate SBS, it is classified as *reserved* to avoid placing another wake-up order for the same SBS (for another UE request). Although this approach decreases the blocking probability of SCN, the energy consumption is likely to increase since sleeping SBSs getting wake-up orders cannot remain in their low-power consumption states.

4.3.5 Evaluation of Computational Complexity of Algorithms

In this section, we investigate the computational complexity of the OOS schemes considered in this work. To this end, we assume a large circular area with a radius of $k_\mathrm{m}R_\mathrm{th}$ such that $A = \pi\,(k_\mathrm{m}R_\mathrm{th})^2$. This area is populated with N_c SBSs and N_u UEs such that and $\rho_\mathrm{c} = \frac{N_\mathrm{c}}{A}$ and $\rho_\mathrm{u} = \frac{N_\mathrm{u}}{A}$. Let Π_I, Π_S, and Π_A are the probability of any SBS being in the *idle*, *sleep*, and *active* modes, respectively. The computational complexity of each OOS scheme can then be given as follows.

4.3.5.1 ROO Algorithm

In the ROO algorithm, the computational complexity is equivalent to the time complexity of generating a random number, which is $\mathcal{O}(1)$.

Algorithm 4 Wake-up Control Based Service Request Handling (WUC)

```
**Input**: UE service request arrival at $t_{\text{now}}$    ▷ $t_{\text{now}}$ is the current time
$t_{\text{deadline}} \leftarrow t_{\text{now}} + w_{\text{t}}$
**while** $t_{\text{now}} \leq t_{\text{deadline}}$ **do**
    $\text{SBS}_{\text{best}} = \text{WUC}(t_{\text{deadline}}, t_{\text{now}}, \mathcal{S}_{\text{all}})$    ▷ $\mathcal{S}_{\text{all}}$ is the set of all SBSs
    update $t_{\text{now}}$
**end while**
**if** $\text{SBS}_{\text{best}} == \varnothing$ **then**
    service request is blocked
**else**
    associate UE to $\text{SBS}_{\text{best}}$
**end if**
**procedure** $\text{WUC}(t_{\text{deadline}}, t)$    ▷ WUC algorithm
    $\mathcal{S} \leftarrow \text{find}_{1 \leq \ell \leq |\mathcal{S}_{\text{all}}|}\,(\,\text{distance}(\mathcal{S}_{\text{all}}(\ell), \text{UE}\,) \leq R_{\text{th}}\,)$
    $\mathcal{S}_{\text{candidate}} \leftarrow \text{find}_{1 \leq \ell \leq |\mathcal{S}|}\,(\,t + \text{bootupTime}(\mathcal{S}(\ell)) \leq t_{\text{deadline}}\,)$
    **if** $\mathcal{S}_{\text{candidate}} == \varnothing$ **then**
        **return** $\varnothing$
    **else**
        $j \leftarrow \text{argmin}_{1 \leq \ell \leq |\mathcal{S}_{\text{candidate}}|}\,\text{distance}(\mathcal{S}_{\text{candidate}}(\ell), \text{UE}\,)$
        **return** $\mathcal{S}_{\text{candidate}}(j)$
    **end if**
**end procedure**
```

4.3.5.2 CLB Algorithm

In the CLB algorithm, the central controller chooses the SBS among the set of all idle SBS, which is $\Pi_\mathrm{I} N_C$. Sorting SBSs based on their loads, and selecting the minimum one has a worst case complexity of $\mathcal{O}((\Pi_\mathrm{I} N_C)^2)$, which is smaller than $\mathcal{O}((1-\Pi_\mathrm{S})^2 N_C^2)$ since $\Pi_\mathrm{I} = 1 - \Pi_\mathrm{S} - \Pi_\mathrm{A} < 1 - \Pi_\mathrm{S}$.

4.3.5.3 LBD Algorithm

In the LBD algorithm, the local breadth-first search (BFS) is bounded either by the maximum search range $k_\mathrm{S} R_\mathrm{th}$, or by the load-based stopping criterion given by (4.35) before reaching the maximum search range. In the worst case scenario, local BFS therefore reaches the maximum search range. The upper bound on the computational complexity can then be found without considering the load-based stopping criterion. Assuming that N_S be the mean number of traversed idle SBSs during the local BFS, we have

$$\begin{aligned} N_S &\leq \rho_\mathrm{c} \Pi_\mathrm{I} (k_\mathrm{S} R_\mathrm{th})^2, && (4.37) \\ &< \rho_\mathrm{c} (1-\Pi_\mathrm{S})(k_\mathrm{S} R_\mathrm{th})^2 = (1-\Pi_\mathrm{S}) \left(\frac{k_s}{k_\mathrm{m}}\right)^2 N_\mathrm{c}, && (4.38) \end{aligned}$$

where (4.37) is formulated based on the observation that the load-based stopping criterion may be satisfied before the maximum search range is reached.

It is well-known that the BFS has a time complexity of $\mathcal{O}(N_\mathrm{S} + E_\mathrm{S})$, where E_S is the number of edges, or, equivalently, the number of neighborhoods between the SBSs. We need the average degree of SBSs, (i.e. average number of SBSs within SBS's communication range) , $\nu_{\mathrm{c}_\mathrm{LBD}}$, to compute E_S. Note that the sum of load factors for each UE is 1 by definition of (4.1). Except UE's with no neighboring SBS, sum of load factors UEs is equal to the sum of load values of cells. If we assume that no SBS is turned off,

$$(1 - e^{-\nu_c}) N_\mathrm{u} \approx \sum_{i=1}^{N_c} L(i), \tag{4.39}$$

where $(1 - e^{-\nu_c})$ is due to the UEs having no SBS in the communication range. If A is very large (4.39) holds with equality. So approximation sign is due to the edge effects. Similarly, if Π_S proportion of the SBSs are turned off randomly

$$(1 - e^{-\nu_c(1-\Pi_\mathrm{S})})N_\mathrm{u} \approx \sum_{i=1}^{N_c(1-\Pi_\mathrm{S})} L(i), \tag{4.40}$$

$$< \sum_{i=1}^{N_c(1-\Pi_\mathrm{S})} L_\mathrm{LBD}(i), \tag{4.41}$$

where (4.41) is due to the fact that the LBD algorithm keeps the SBSs having larger load values in idle mode. We therefore observe that the only way for the approximation to hold is to increase ν_c at the left side of (4.40). The average SBS degree therefore increases with LBD. and $\nu_{\mathrm{c}_\mathrm{LBD}}$ satisfies

$$(1 - \Pi_\mathrm{S})\nu_\mathrm{c} \leq \nu_{\mathrm{c}_\mathrm{LBD}} \leq \nu_\mathrm{c}. \tag{4.42}$$

A bound on E_S can then be obtained as

$$E_\mathrm{S} = \frac{1}{2}\sum_{i=1}^{N_\mathrm{S}} \text{degree (i)} \tag{4.43}$$

$$= \mathbb{E}[N_\mathrm{S}]\mathbb{E}[\text{SBS degree}] \tag{4.44}$$

$$< \frac{1}{2}\nu_{\mathrm{c}_\mathrm{LBD}}(1 - \Pi_\mathrm{S})\left(\frac{k_s}{k_\mathrm{m}}\right)^2 N_\mathrm{c} \tag{4.45}$$

$$< \frac{1}{2}\nu_\mathrm{c}(1 - \Pi_\mathrm{S})\left(\frac{k_s}{k_\mathrm{m}}\right)^2 N_\mathrm{c} \tag{4.46}$$

$$= \frac{1}{2}(1 - \Pi_\mathrm{S})\frac{k_s^2}{k_\mathrm{m}^4}N_\mathrm{c}^2, \tag{4.47}$$

where coefficient $\frac{1}{2}$ in (4.43) is due to counting each SBS twice for single SBS neighborhood. (4.44) is due to independent locations of SBSs in HPPP. (4.45) follows from (4.38). Finally, (4.46) is due to (4.42).

SBS reaches average of $\nu_c\Pi_\mathrm{I}$ SBSs immediately, which has mean sorting complexity $(\nu_c\Pi_\mathrm{I})^2$. If stopping condition (4.35) is not satisfied, further SBSs reached by range expansion will be added.

Since the initial sorting is done, adding a new load value to a sorted array has linear complexity with array size. Besides, in our algorithm, maximum search range can expanded at most by R_{th} at one time. If current search range is $(i-1)R_{\text{th}}$, and algorithm is reaching new idle SBSs' in iR_{th} range, expected sorting complexity $\mathcal{O}(S_{LBD}$ can be written as

$$\mathcal{O}(S_{LBD}) = (\nu_c \Pi_{\text{I}})^2 + \sum_{i=2}^{k_{\text{S}}} \sum_{j=0}^{\infty} \sum_{t=1}^{\infty} \sum_{m=0}^{t-1} (j+m) p_c(j) p_c(t) \tag{4.48}$$

$$= (\nu_c \Pi_{\text{I}})^2 + \sum_{i=2}^{k_{\text{S}}} \sum_{j=0}^{\infty} \sum_{t=1}^{\infty} \left(jt + \frac{t(t-1)}{2} \right) p_c(j) p_c(t)$$

$$= (\nu_c \Pi_{\text{I}})^2 + \sum_{i=2}^{k_{\text{S}}} \Big[\sum_{j=0}^{\infty} \sum_{t=0}^{\infty} jt p_c(j) p_c(t) + \sum_{j=0}^{\infty} p_{\text{c}}(j) \sum_{t=0}^{\infty} \frac{1}{2} t(t-1) p_{\text{c}}(t) \Big] \tag{4.49}$$

where $p_c(j)$ is Poisson with mean $(i-1)^2 \Pi_{\text{I}} \nu_c$. Since between radius $(i-1)R_{\text{th}}$ and iR_{th}, $(2i-1)\Pi_{\text{I}}\nu_c$, there are $(i^2-(i-1)^2)\nu_c \Pi_{\text{I}}$ SBSs, $p_c(t)$ is Poisson with mean $(2i-1)\nu_c \Pi_{\text{I}}$. $p_c(j)$ and $p_c(t)$ are independent due to disjoint areas. Then, (4.49) can be rewritten as

$$\mathcal{O}(S_{LBD}) = (\nu_c \Pi_{\text{I}})^2 + \sum_{i=2}^{k_{\text{S}}} \Big[(2i^3 - i^2)(\Pi_{\text{I}} \nu_c)^2 + \frac{1}{2}((2i-1)^2 (\nu_c \Pi_{\text{I}})^2) \Big] \tag{4.50}$$

$$= (\nu_c \Pi_{\text{I}})^2 + (\nu_c \Pi_{\text{I}})^2 \sum_{i=2}^{k_{\text{S}}} (2i^3 + i^2 - 2i + \frac{1}{2})$$

$$\approx \mathcal{O}(k_{\text{S}}^4 (\nu_c \Pi_{\text{I}})^2) \tag{4.51}$$

$$= \mathcal{O}(k_{\text{S}}^4 \Pi_{\text{I}}^2 \frac{N_c^2}{k_{\text{m}}^4}) < \mathcal{O}((1-\Pi_{\text{S}})^2 \left(\frac{k_{\text{S}}}{k_{\text{m}}} \right)^4 N_c^2)$$

Note that the DLB algorithm has a reasonable complexity for $\frac{k_s^4}{k_{\text{m}}^4} \ll 1$ since it becomes more likely to find an SBS with sufficiently low load value without having to search through the entire SCN. Note that we our complexity analysis considers worst case bounds for sorting and inserting. For example, if central controller keeps a pre-sorted list of load values, finding SBS with minimum load value has $log\big((1-\Pi_{\text{S}})N_C\big)$. Investigating tighter bounds is left as a future work.

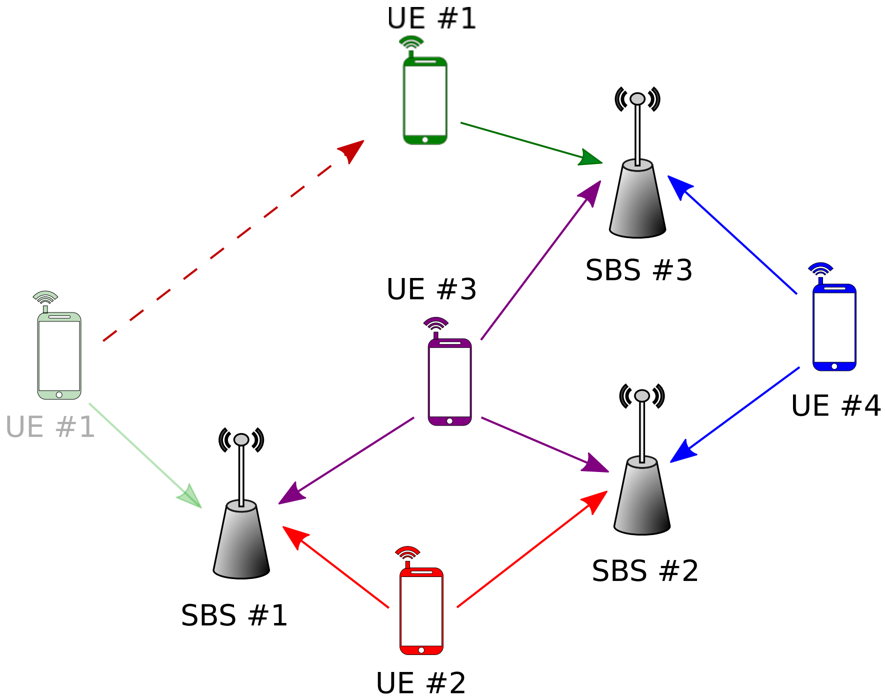


Figure 4.5: Updating load values when UEs move

4.3.5.4 WUC Algorithm

In the WUC algorithm, central controller wakes up an SBS within the communication range of UE, which takes $\mathcal{O}(\Pi_S \nu_c) = \mathcal{O}(\Pi_S \frac{N_c}{k_m^2})$ steps.

4.4 Adaptation of Load Based Algorithms to User Mobility

Load values can be effectively updated in distributed manner If UE moves from one location to another, SBS can recognize that UE moved out of its range by controlling uplink control signals. Similarly, UE can update its new load factor by processing downlink signals from SBS whitin its communication range. In Figure 4.5, UE$_1$ moves out of the range of SBS$_1$. Load factor of UE does not change. $w_1 = w_{1'} = 1$, $L_1 = L_{1'} - w_{1'} = \frac{11}{6} - 1 = \frac{5}{6}$, and $L_3 = L_{3'} + w_1 = \frac{5}{6} + 1 = \frac{11}{6}$. We use a prime symbol to represent prior values of load and load factors. Load values can be updated without additional signaling.

In case of mobility, centralized load based algorithm can work without change. However, depending on the mobility model, distributed load based algorithm may or may not work. If UEs move randomly, Poisson property still hold, so distribution of load can be obtained. However, if the mobility model has specific rules aimed to represent human behavior, stationarity of Poisson process may no longer hold. In this case, analyzing load distribution becomes challanging.

4.5 Simulation Results

In this section, we present numerical results for the performance of proposed load definition in representing the actual traffic load of SCN, and novel load based OOS strategies. In particular, performance of the novel CLB and DLB algorithms are evaluated in comparison to the ROO and WUC algorithms as the benchmark OOS strategies, and the static topology without dynamic OOS approach.

We assume a circular area with a radius of 250 m for the deployment of UEs and SBSs, and the results are averaged over $1,000$ iterations and $10,000$ seconds of simulation time. In terms of overall SCN traffic, we consider two main scenarios: low network utilization (1%) and (relatively) high network utilization (20%). Rationale behind the low and relatively high network utization is to leave enough room to effectively apply OOS strategies. For utilization levels above 20%, more and more SBSs would be occupied all the time. Turning off SBSs at high utilization levels is detrimental for QoS. For delayed access scheme, we assume a sufficiently large but reasonable UE delay tolerance of 60 sec (as well as zero tolerable delay), which enables WUC algorithm to attain its best performance, and, hence, the performance gap between WUC and other strategies becomes apparent. All the simulation parameters for low and high utilization scenarios are listed in Table 4.4.

4.5.1 Performance Metrics

In the performance analysis, we consider the following criteria.

- *Blocking probability*: The fraction of rejected service requests among all, which is basically due to sleeping or occupied (i.e., actively transmitting) SBSs, which is given as

$$\mathsf{P}_{\mathsf{block}} = \frac{\text{number of rejected service requests}}{\text{total number of service requests}}. \tag{4.52}$$

- *Average throughput*: The total number of bits transmitted averaged over the total simulation time, which is also normalized with respect to the number of users as follows

$$\mathsf{R}_{\mathsf{SCN}} = \frac{\text{total number of transmitted bits}}{\text{number of users} \times \text{simulation time}} \text{ (bps).} \tag{4.53}$$

The number of transmitted bits in (4.53) is given by the Shannon capacity formula as follows

$$\mathsf{R} = \mathsf{BW} \log_2 (1 + \mathsf{SINR}), \tag{4.54}$$

where BW is the transmission bandwidth, and SINR is the signal-to-interference-plus-noise ratio. We assumed frequency reuse in SCN network. Bandwidth is used to compute the throughput of UE. If the allocated bandwidth is large, our algorithms can still operate without any change. Assuming the association between ith UE and jth SBS, the respective SINR at the UE side is defined as follows

$$\mathsf{SINR}_{ij} = \frac{d_{ij}^{-\alpha}}{\sum\limits_{\ell \neq j} d_{i\ell}^{-\alpha} + 1/\mathsf{SNR}}, \tag{4.55}$$

where d_{ij} is the distance between ith UE and jth active SBS, α is the path loss (PL) exponent, and SNR is the signal-to-noise ratio.

- *Normalized Energy Efficiency*: The amount of energy consumed for each transmitted bit averaged over the total simulation time, which is also normalized by the number of users and the maximum power $\mathsf{P}_{\mathsf{max}}$ associated with the active state.

$$\mathsf{EE} = \frac{\mathsf{R}_{\mathsf{SCN}}}{\text{total energy consumption}} \times \mathsf{P}_{\mathsf{max}}\,(\text{bps/joule}). \tag{4.56}$$

Note that the power consumption of an SBS at each state is given in Table 4.1 as the fraction of the maximum power $\mathsf{P}_{\mathsf{max}}$, and we therefore use these power levels while computing (4.56).

4.5.2 Load Distribution Verification

In Figure 4.6, we depict the CDF and PDF of the SCN traffic load for range-dependent UE densities of $\nu_{\mathrm{u}} = \{3, 5, 10\}$, where extensive simulation results are provided along with the analytical results computed using (4.33). We observe that the analytical results nicely match the simulations for all three UE densities, which verifies the respective derivation in Section 4.2. Accuracy of approximate distribution depends on achievable precision of load values which ultimately depends the precision of load factor. Precision of load improves as load factor can take small values. When number of UE around SBS is small, load factor is high. Therefore, as we increase UE's range, the load factor of UE becomes smaller, and the approximate load distribution converges to the exact load distribution. We, therefore, start observing approximate load distribution matches with simulation results as the average number of UEs around cell increases.

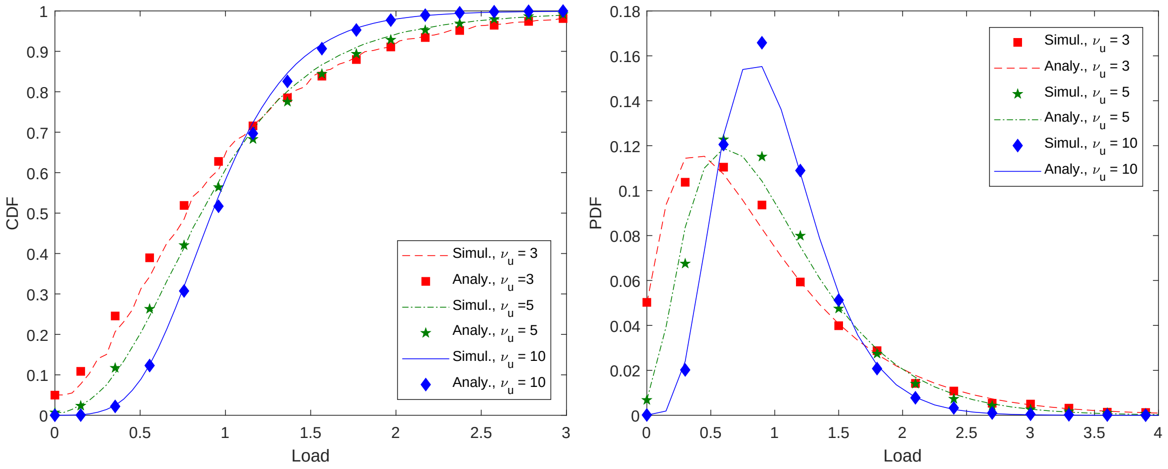
In Figure 4.7, the simulation results shows the load distribution under erroneous location measurements. Standard deviation of error is $\sigma = R_{\mathrm{th}} \frac{\%\text{ of error}}{100}$. We observe that the load distribution does not change by Gaussian error.

4.5.3 Low Utilization Performance

In this subsection, we consider the performance of OOS strategies under a low network utilization scenario, where the UE service request rate and average file size are $1/\lambda_{\mathrm{U}} = 1000\,\mathrm{s}$ and $1/\lambda_{\mathrm{F}} = 1\,\mathrm{MB}$, respectively. Together with the UE and SBS densities given in Table 4.4. Respective

Table 4.4: Simulation Parameters

Parameter	Value
Cell density (ρ_{c})	0.0005 m^{-2}
User density (ρ_{u})	0.0005 m^{-2}
Service request rate (λ_{U})	$\{0.001, 0.01\}$ s^{-1}
Average file size ($1/\lambda_{\mathrm{F}}$)	$\{1, 2\}$ MB
Sleep rate (λ_{S})	$\{0.001, 0.002\}$ s^{-1}
Tolerable delay (w_{t})	$\{0, 60\}$ s
Threshold distance (R_{th})	50 m
Bandwidth (BW)	1 MHz
Signal-to-Noise Ratio (SNR)	20 dB
Threshold probability for DLB (κ)	0.3
Maximum search range for DLB	$3 \times R_{\mathrm{th}}$
Path loss exponent (α)	2



(a) CDF

(b) PDF

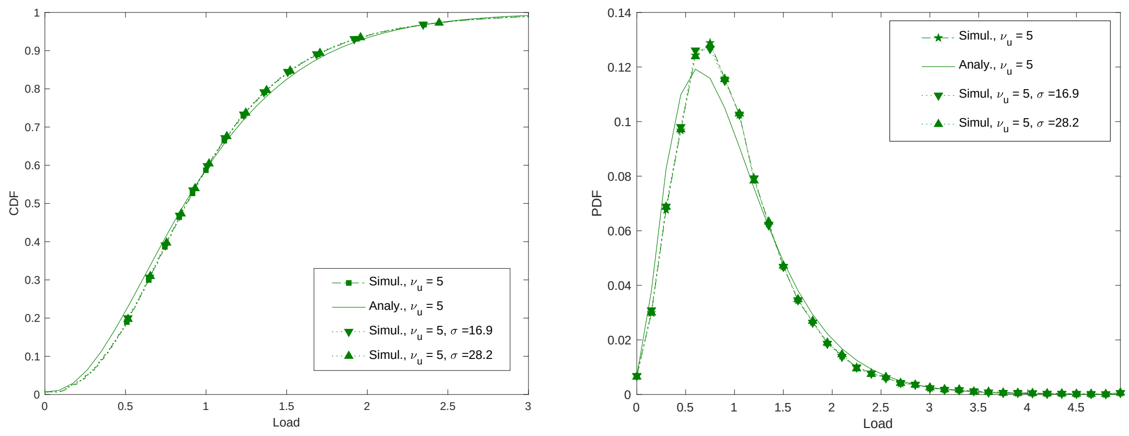
Figure 4.6: Analytical and simulation results for load distribution for range-dependent UE densities of $\nu_{\mathrm{u}} = \{3, 5, 10\}$ and $\rho_{\mathrm{c}}/\rho_{\mathrm{u}} = 1$.

network utilization is on the order of 1% based on the utilization results in Figure 4.8.

In Figure 4.9, we present blocking probability results for all the algorithms under consideration against varying *on-ratio* (i.e. fraction of non-sleeping SBSs). In particular, we take into account the effect of sleep rate λ_S (or equivalently sleep period $1/\lambda_\mathrm{S}$) and waiting time w_t by assuming $1/\lambda_\mathrm{S} \in \{500\,\mathrm{s}, 1000\,\mathrm{s}, \infty\}$ and $w_\mathrm{t} \in \{0, 60\,\mathrm{s}\}$. Note that $1/\lambda_\mathrm{S} \to \infty$ corresponds to a scenario with no dynamic OOS events, i.e., topology of non-sleeping SBSs does not change once it is initialized at the beginning. We therefore describe the respective load based algorithm simply with LB since either centralized or distributed strategy (i.e., in CLB and DLB) is only applicable with dynamic on/off events occurring after initialization.

We observe in Figure 4.9 that the blocking probabilities for any OOS algorithm decrease as either more SBSs become available (i.e., increasing on ratio), or tolerable delay gets larger (i.e., more room to meet UE service requests). In particular, the load based CLB and DLB perform much better than the random scheme ROO in terms of achieving less blocking events (i.e., rejected UE requests). Note that CLB and DLB actually have the same performance for any choice of on ratio, and we therefore referred to this common performance as CLB/DLB. This equality underscores the power of DLB especially for large-scale SCNs in the sense that DLB does not need information of *all* SBSs (i.e., in contrast to CLB) to decide the next SBS to turn off, and is hence more efficient to implement. Considering a wide range of reasonable non-sleeping SBS fractions (i.e., greater than 0.5 for a realistic SCN), CLB/DLB is shown to attain the performance of more complex WUC scheme, where ROO still falls short of that level. For 1% network utilization, load based algorithms still have an on-ratio of 0.5. On the other hand, WUC allows 0.1 on ratio for same blocking probability . It's clear that optimization of sleep time may ensure much higher energy saving. While maintaining same QoS it may be possible to design more sophisticated algorithms considering both sleep time, and load to achieve further energy savings.

In Figure 4.9, the response of random and load based algorithms to the choice of sleep period $1/\lambda_\mathrm{S}$ and waiting time w_t are observed to have some interesting differences. Assuming zero tolerable delay (i.e., $w_\mathrm{t} = 0$), the blocking probability of random scheme ROO does not change at all



(a) CDF

(b) PDF

Figure 4.7: Analytical and simulation results for load distribution for range dependent UE densities of $\nu_u = 5$ and $\rho_c/\rho_u = 1$. R_{th} = 56.4. Localization error(%) = {0, 30, 50} σ = {16.9, 28.2}.

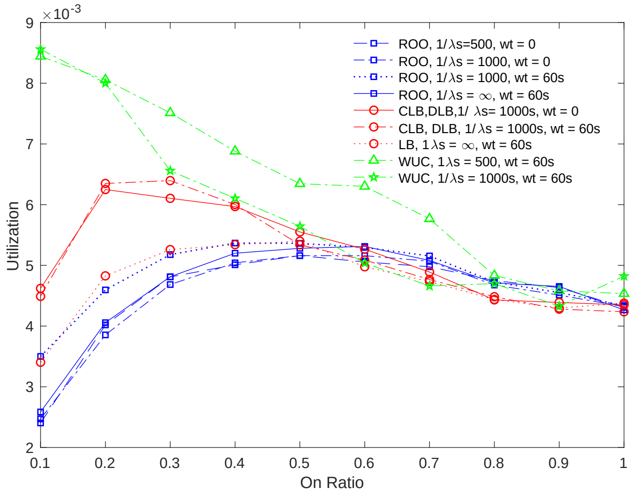


Figure 4.8: Network utilization along with on ratio (i.e., fraction of non-sleeping SBSs) for $1/\lambda_S \in \{500\,s, 1000\,s, \infty\}$ and $w_t \in \{0, 60\,s\}$. Low network utilization of 1% (i.e., $1/\lambda_U = 1000\,s$ and $1/\lambda_F = 1\,MB$).

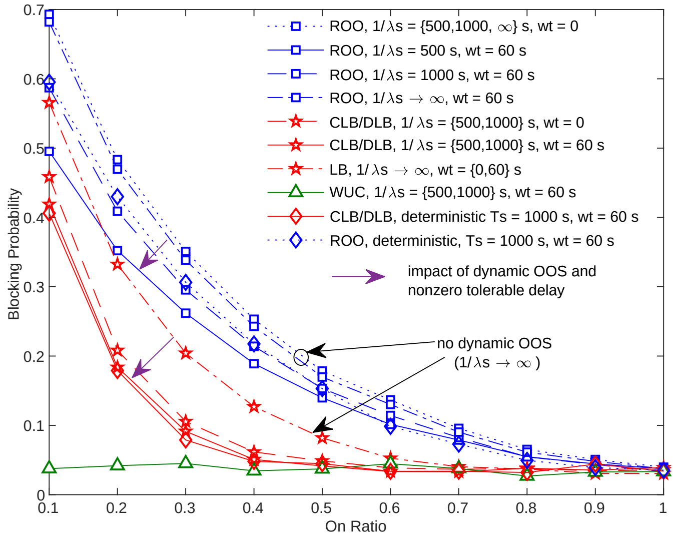
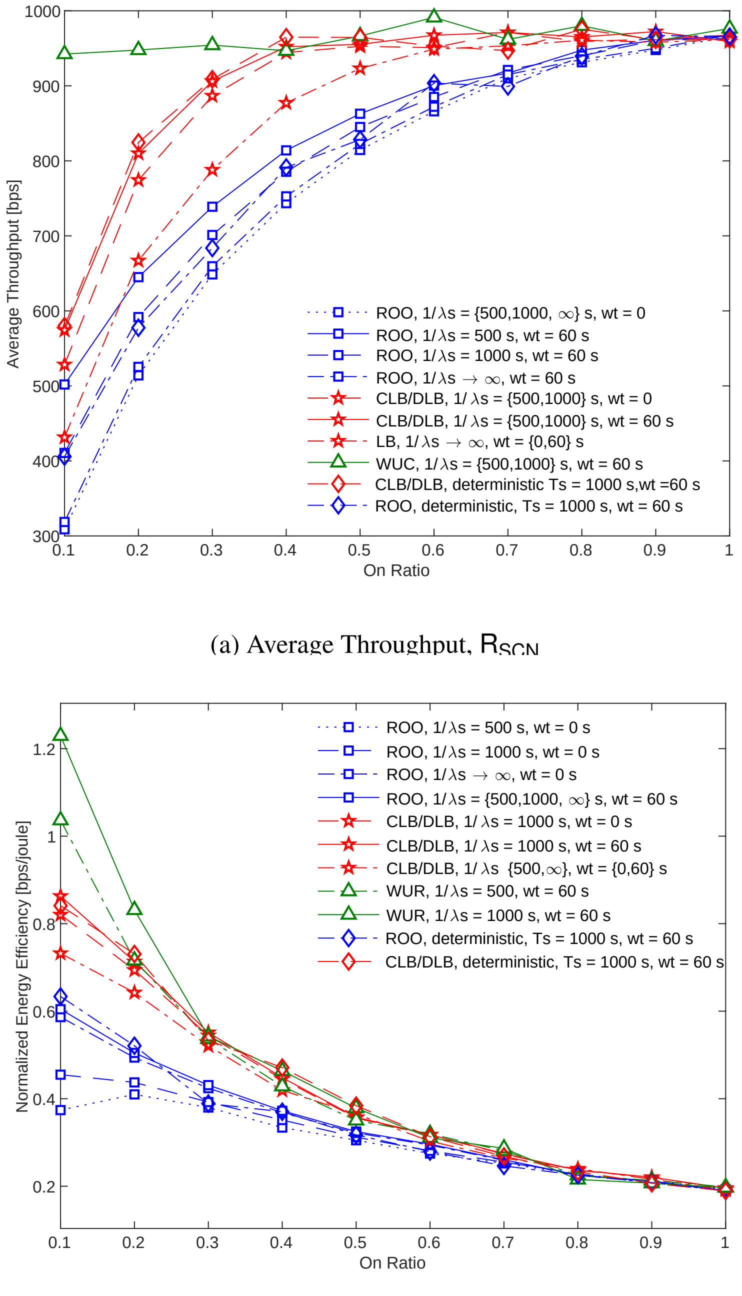


Figure 4.9: Blocking probability $\mathsf{P}_{\mathsf{block}}$ along with on ratio (i.e., fraction of non-sleeping SBSs) for $1/\lambda_{\mathrm{S}} \in \{500\,\mathrm{s}, 1000\,\mathrm{s}, \infty\}$ and $w_{\mathrm{t}} \in \{0, 60\,\mathrm{s}\}$ assuming low network utilization of 1% (i.e., $1/\lambda_{\mathrm{U}} = 1000\,\mathrm{s}$ and $1/\lambda_{\mathrm{F}} = 1\,\mathrm{MB}$). Term *no dynamic* indicates that either random or load based decision is initially made, and SBS do not turn on and off

along with $1/\lambda_{\mathrm{S}}$ even considering the no dynamic OOS case (i.e., $1/\lambda_{\mathrm{S}} \rightarrow \infty$). When we consider nonzero tolerable delay (i.e., $w_{\mathrm{t}} = 60\,\mathrm{s}$), we start observing significant performance improvement in ROO along with decreasing $1/\lambda_{\mathrm{S}}$, where the best performance occurs at $1/\lambda_{\mathrm{S}} = 500\,\mathrm{s}$. On the other hand, load based CLB/DLB achieves significantly better performance for $1/\lambda_{\mathrm{S}} = \{500\,\mathrm{s}, 1000\,\mathrm{s}\}$ (as compared to no dynamic OOS case) even under zero tolerable delay condition. When a nonzero tolerable delay (i.e., $w_{\mathrm{t}} = 60$ s) is further assumed, the best performance is even superior to that of the zero tolerable delay, but the respective performance gap remains marginal. As a result, CLB/DLB is more robust to *delay intolerance* while random scheme ROO requires *longer tolerable delays* for performance improvement. In addition, applying OOS dynamically is useful for ROO only when the delay tolerance is sufficiently large. If delay budget with respect to sleeping time is too small, UE will not be able to take advantage of delay. On the other hand, dynamic OOS improves performance of CLB/DLB in both delay tolerant and intolerant SCNs.

In Figure 4.10, we present the respective network throughput and normalized energy efficiency results. We observe that the network throughput performance in Figure 4.10(a) shows closely



(a) Average Throughput, $\mathsf{R}_{\mathsf{SCN}}$

(b) Normalized Energy Efficiency, EE

Figure 4.10: Average throughput and normalized energy efficiency along with on ratio (i.e., fraction of non-sleeping SBSs) for $1/\lambda_{\mathrm{S}} \in \{500\,\mathrm{s}, 1000\,\mathrm{s}, \infty\}$ and $w_{\mathrm{t}} \in \{0, 60\,\mathrm{s}\}$ assuming low network utilization of 1% (i.e., $1/\lambda_{\mathrm{U}} = 1000\,\mathrm{s}$ and $1/\lambda_{\mathrm{F}} = 1\,\mathrm{MB}$).

related behavior to the blocking probability results (i.e., network throughput increases with decreasing blocking probability, and vice versa). In particular, we observe no significant average throughput loss when as many as 40% of SBSs are in sleeping states. On the other hand, the average throughput of ROO keeps decreasing continuously as more SBSs are put into sleeping states, which finally reads as high as 20% throughput loss for non-sleeping SBSs fraction of 40%.

The normalized energy efficiency results in Figure 4.10(b) involve some interesting conclusions as follows. 1) Energy efficiency of ROO is worse than that of CLB/DLB whereas CLB/DLB is as energy-efficient as the more complex WUC scheme for non-sleeping SBS fractions greater than 30%. 2) Although ROO attains the maximum throughput only under nonzero tolerable delay (see Figure 4.10(a)), the maximum energy efficiency can be achieved under both zero and nonzero tolerable delays. In particular, while the maximum energy efficiency of ROO is invariant to sleep period under nonzero tolerable delay, the best sleep period turns out to be $\lambda_S \to \infty$ under zero tolerable delay. As a result, *the energy efficiency for ROO under zero tolerable delay gets maximized when OOS scheme is not applied dynamically* (i.e., no on/off events after initialization). 3) Although network throughput for CLB/DLB is maximized for $1/\lambda_S \in \{500\,\mathrm{s}, 1000\,\mathrm{s}\}$ with a significant gap between the no dynamic OOS case (i.e., $\lambda_S \to \infty$), the energy efficiency gets maximized only for $1/\lambda_S = 1000\,\mathrm{s}$ under any choice of tolerable delay. *Regardless of the particular tolerable delay in CLB/DLB, assigning short sleep time is therefore as energy inefficient as keeping SBSs in sleep states for very long*, which identifies an optimal sleep period in between.

In Figures 4.9, and 4.10, deterministic sleep periods are compared with exponential sleep periods. In case of having a deterministic sleep period for each cell, the sleeping cell will be in one of these sleep modes with probability one according to our simple rule in Table 4.2. On the contrary, random sleep periods ensure that cell can be in any sleeping mode in various sleep lengths with non-zero probability. So, selection of the random period does not bias results in favor of load based algorithms, it actually demonstrates that load based algorithms can efficiently handle varying sleep periods, and sleep modes while decreasing overall network power consumption. In other words, instead of evaluating the performance of load based algorithms at fix deterministic periods, introducing randomness into sleep periods enables testing robustness of the algorithms in a large set of sleeping periods.

4.5.4 High Utilization Performance

We now consider a high network utilization scenario with the UE service request rate of $1/\lambda_{\mathrm{U}} = 100\,\mathrm{s}$ and the average file size of $1/\lambda_{\mathrm{F}} = 1\,\mathrm{MB}$. The respective utilization is on the order of 20% as shown in Figure 4.11 We assume a representative finite sleep time period together with no dynamic OOS case, i.e., $1/\lambda_{\mathrm{S}} \in \{1000\,\mathrm{s}, \infty\}$, together with both zero and nonzero tolerable delays, i.e., $w_{\mathrm{t}} \in \{0, 60\,\mathrm{s}\}$. In Figure 4.9, we present blocking probability results along with on ratio. As before, we observe that the performances of CLB and DLB are much better than that of ROO, and are the same as that of WUC whenever at least 50% of the SBSs are non-sleeping. In addition, performane of DLB has a close to that of CLB, as before. We also observe that the performance of any OOS algorithm improves together with either nonzero tolerable delay, or applying dynamic OOS (i.e., $1/\lambda_{\mathrm{S}} = 1000\,\mathrm{s}$ instead of $1/\lambda_{\mathrm{S}} \to \infty$) on top of that. Regardless of the particular OOS strategy, the blocking probabilities are observed to be higher than those in Figure 4.9 as the fraction of non-sleeping SBSs decreases, which is due to the increased network utilization.

In Figure 4.13, we demonstrate the average throughput and normalized energy efficiency performances against on ratio. As before, the average throughput results in Figure 4.13(a) indicate that the performance of CLB and DLB are much better than that of ROO, and are the same as WUC for a broad range of non-sleeping SBS fractions (i.e., greater than 0.5). In particular, the average throughput of either CLB or DLB remains almost unchanged even when 50% of the SBSs are put into sleeping states, while the respective loss in ROO throughput appears to be between 10%-30% for the same on ratio. Note that the average throughput results in Figure 4.13(a) are much higher as compared to that of Figure 4.10(a) owing to the increased network utilization. In addition, the average throughput increases for all the OOS algorithms as UEs become more delay tolerant.

We also present the respective normalized energy efficiency results in Figure 4.13(b) for this high utilization scenario. We observe that the energy efficiency of CLB and DLB gets maximized

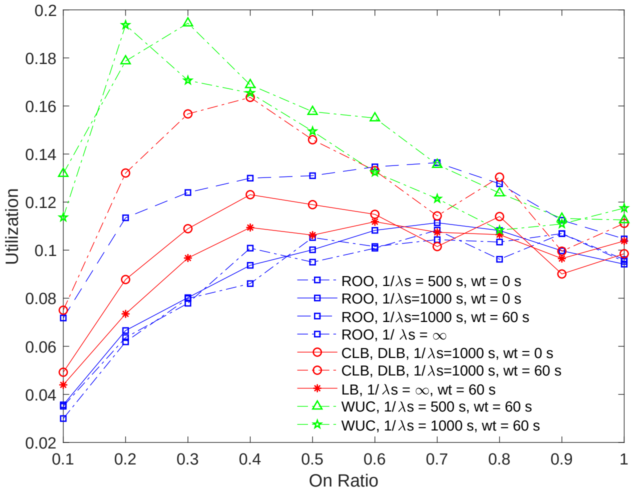


Figure 4.11: Network utilization along with on ratio (i.e., fraction of non-sleeping SBSs) for $1/\lambda_S \in \{500\,\mathrm{s}, 1000\,\mathrm{s}, \infty\}$ and $w_t \in \{0, 60\,\mathrm{s}\}$. High network utilization of 20% (i.e., $1/\lambda_U = 1000\,\mathrm{s}$ and $1/\lambda_F = 1\,\mathrm{MB}$).

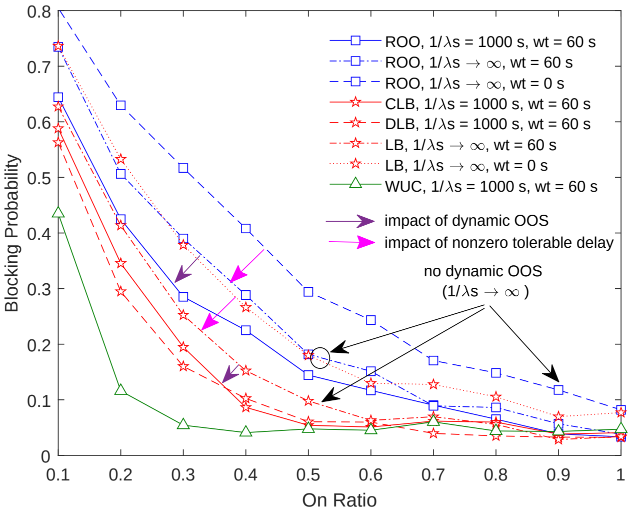
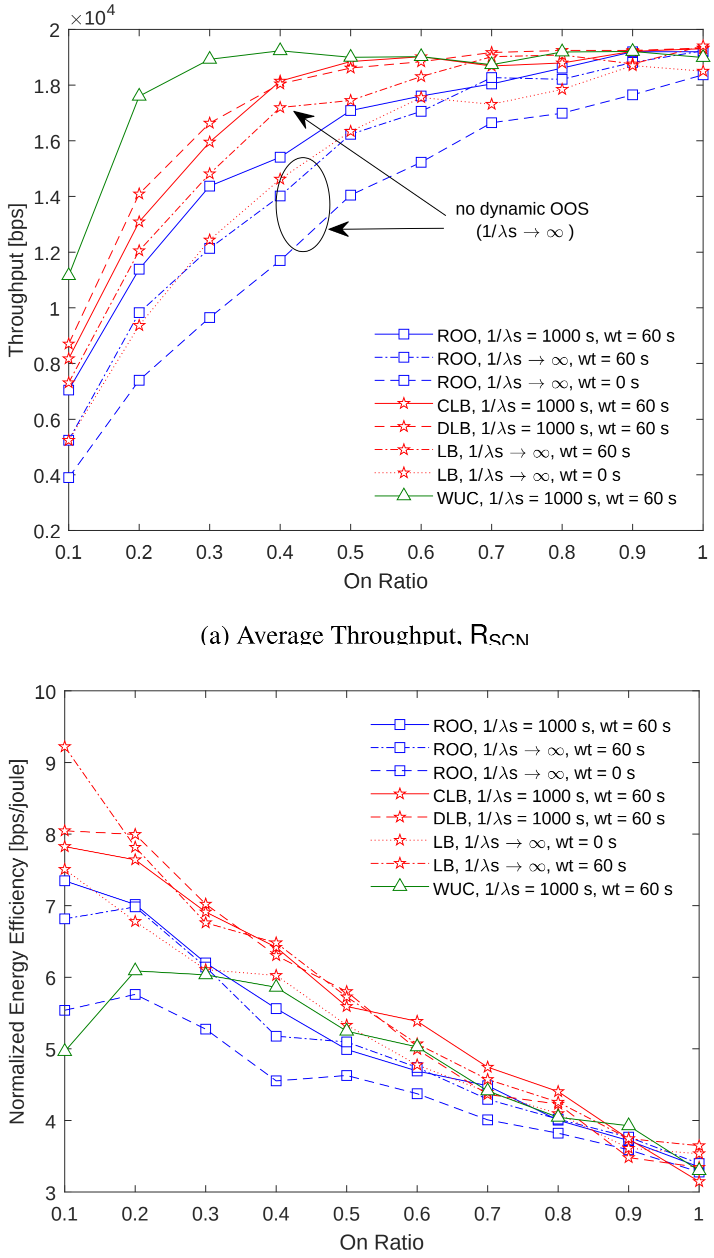


Figure 4.12: Blocking probability $\mathsf{P}_{\mathsf{block}}$ along with on ratio (i.e., fraction of non-sleeping SBSs) for $1/\lambda_S \in \{1000\,\mathrm{s}, \infty\}$ and $w_t \in \{0, 60\,\mathrm{s}\}$ assuming low network utilization of 20% (i.e., $1/\lambda_U = 100\,\mathrm{s}$ and $1/\lambda_F = 2\,\mathrm{MB}$).



(a) Average Throughput, $\mathsf{R}_{\mathsf{SCN}}$

(b) Normalized Energy Efficiency, EE

Figure 4.13: Average throughput and normalized energy efficiency along with on ratio (i.e., fraction of non-sleeping SBSs) for $1/\lambda_{\mathrm{S}} \in \{1000\,\mathrm{s}, \infty\}$ and $w_{\mathrm{t}} \in \{0, 60\,\mathrm{s}\}$ assuming low network utilization of 20% (i.e., $1/\lambda_{\mathrm{U}} = 100\,\mathrm{s}$ and $1/\lambda_{\mathrm{F}} = 2\,\mathrm{MB}$).

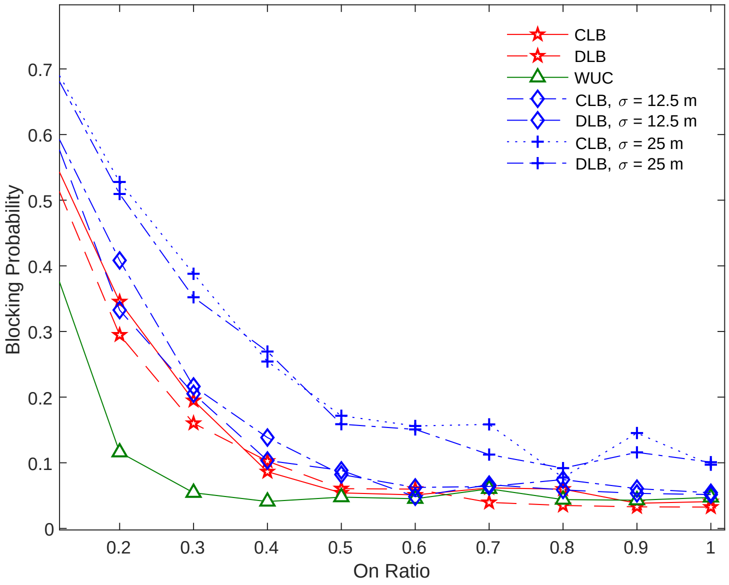


Figure 4.14: Effect of localization error in SCN along with on ratio (i.e., fraction of non-sleeping SBSs) for $1/\lambda_{\mathrm{S}}$ = 1000 s and w_{t} = 60 s assuming high utilization of $\approx 20\%$ (i.e., $1/\lambda_{\mathrm{U}} = 1000$ s and $1/\lambda_{\mathrm{F}} = 2$ MB), R_{th} = 50, $\sigma = \{12.5, 25\}$.

with the nonzero tolerable delay (i.e., $w_{\mathrm{t}} \in 60$ s), which is superior to not only ROO but also more sophisticated WUC scheme. This interesting result indicates that although the average network throughput is maximized (through decreasing blocking probabilities) by the deliberate wake-up control mechanism of WUC, the resulting scheme becomes less energy-efficient. In other words, while the network rejects a smaller number of UE requests by further incorporating the sleeping SBSs, the overall network starts consuming more power since not all SBSs are allowed to complete their full sleep period. As a result, the energy efficiency of WUC deteriorates and falls even below ROO under certain settings. We therefore conclude that, *in contrast to low utilization scenario, the energy efficiency of WUC can be poor under high network utilization, although the associated average throughput might still be the best.*

Figure 4.14 shows that load based algorithms still save energy but increasing localization error cause UEs to consider further cells closer, which decreases the throughput, increases average session period, increases cell utilization. Due to this proportion of blocked service requests increase.

4.6 Conclusion

In this chapter, we considered OOS strategies to implement energy-efficient SCNs. In particular, we propose a novel load definition for the SCN traffic, and derived its approximate distribution rigorously. Two novel load based OOS algorithms (CLB and DLB) are also proposed together with two benchmark strategies ROO as a simple baseline and WUC as the most sophisticated. We show that CLB and DLB perform better than ROO, and have similar performance as compared to WUC under low traffic periods when proportion of on SBS are around 50%. We also observe that when the proportion of on SBSs is below 0.2, WUC performs significantly better. As a future work, it is necessary to design energy saving algorithms that take into consideration of not only traffic load but also sleep time based on number UEs around SBS. We finally show that the performance of CLB can be efficiently attained by DLB in a distributed fashion relying on the statistics of the traffic load. As a future work, traffic load model can be extended to capture diverse mobile usage patterns. Besides, wake-up control and load based schemes can be extended by considering mobile power consumption, and macrocell serving capacity in mid-traffic and high-traffic profiles. Moreover, algorithms determining optimal sleep state and sleep duration based on load metric can be developed.

Chapter 5: Capacity and Energy-Efficiency Analysis of Delayed Access Scheme

The advent of small base stations (SBSs) offers not only opportunities in terms of high capacity and flexibility but also challenges. One important challange is their irregular topology. Therefore, it requires new methods to accurately analyze, signal to interference ratio (SIR), and signal to interfrence plus noise ratio (SINR). Despite that the Wyner model has been widely adopted to analyze SINR, Xu et al. [108] argues that accuracy of the Wyner model highly depends on the number of interferers. If there is a large number of interferers, model is precise, if not, the Wyner is quite inacurate.

In this chapter, we made use of the assumption of random SBS distribution from a different perspective to analyze delayed access scheme. if UE can delay its transmission time, it may have the opportunity to decrease the distance to the SBS, which can be translated as both higher SINR and bitrate. Especially in large scale SBS deployments, it is worth modeling relationship between UE's access delay and SINR level and similar key metrics.

There is a rich literature on the analysis of SINR distribution in random networks using stochastic geometry. Andrews et al. derived [74] tractable expressions of coverage probability and bit-rate of a user in a random location. In [75], control signals are muted to improve SINR and save power in an adaptive manner. In [109], bit-rate and energy efficiency (EE) analysis of random cellular networks are given. Then, EE maximization algorithms are offered.

In Chapter 3 of this thesis, we introduced a simple access scheme. A slightly different access scheme is also adopted in Chapter 4 (see Figure 4.1 and Figure 3.2). In Chapter 3, it is shown that advantage of access delay becomes prominent in small cell networks operating under energy-saving policy since access delay gives UE the opportunity to receive service from a closeby cell in an energy-saving mode instead of connecting a distant SBS. We showed in Chapter 4 that depending on the length of UE's tolerable delay, a nearby SBS in sleep mode can be awakened and blocking probability can be reduced significantly.

Despite the useful findings regarding the access delay, the energy-delay, and capacity-delay tradeoff inherent in this asccess scheme is not well investigated. So, we bring attention to the effect of delayed access scheme on key performance metrics of small cells. We analyze energy-efficiency improvement by delaying the user equipment's (UE) access to a SBS. It is shown how UE's delayed access strategy affect SINR distribution. Obtaining the trade-off between access delay and bit-rate may be of importance for emerging ultra-reliable and low-latency communications (URLLC) [110, 111].

The findings in this chapter are partly published in [112]. Contributions of this chapter are: 1) Expression of coverage probability and transmit rate of a user are derived as a function of transmit range and delay tolerance; 2) For the delayed access scheme, it is shown that there exists optimal threshold distance maximizing coverage probability for target SINR; 3) An efficient numerical algorithm is developed that optimizes bit-rate by computing the optimal threshold distance for given tolerable delay; 4) It is shown that considerable improvement in energy-efficiency of small cell network can be achieved by the delayed access scheme. Besides the theoretical work, it is believed that analyzing the access delay vs SINR distribution gives insight into the design of protocols for delay tolerant applications.

Section 5.1 of this chapter gives a quick recap of system model with an emphasis on its differences from prior models. In Section 5.2, we analyze the distributions of coverage probability of UE as a function of predefined range and delay. In Section 5.3, we derive average achievable bit-rate in UE's access scheme. Section 5.4 derives optimal transmit range that maximizes the coverage probability, and suggest a numerical algorithm to maximize average bit-rate for given delay budget. In Section 5.5, we briefly discuss energy-efficiency improvement by comparing a UE's connection with and without access delay. Section 5.6 provides numerical results that validate our analysis and Section 5.7 concludes the chapter.

5.1 Network Model

In this section, the downlink model of the small cell network is described. Since the network model is very similar to the models used in previous chapters, we prefer highlighting differences. At any given time probability of sleep, idle and active cells are Π_{S}, Π_{I}, and Π_{A}, respectively. In order to preserve ergodicity of the network, we assume these probabilities do not change. In the active mode, an SBS gives data service to one UE with exponential service time with mean $1/\mu$. Similarly, duration of the sleep mode is also exponential with λ_{S}. Once sleep period expires, a sleeping cell becomes idle and avaible for service. Similarly, when the active period of cell finishes, it becomes idle. By knowing the probabilities of the transitions from sleep to idle, and active to idle for individual cell, it is possible to model the postive effect of delay in terms of SINR.

We assume that from UE's pespective, proportion of average idle, active, and sleep SBSs are fixed. UE "sees" fixed proportion of active, idle, and sleep SBSs. In other words, from UE's perspective, SCN is ergodic. So, after a SBS completes its service for UE, it moves to the idle mode, and idle cell that is randomly chosen switches to active mode. UE measures received signal strength when UE decides to initiate connection according to delayed access scheme. If delay budget is very large with respect to sleep time, and range of UE's small, active SBS's switch to idle mode, proportion of active SBSs will decrease. Hence, UE will experience very high SINR.

In practice, it is clear that the random distribution of sleep cell is not the best choice as shown in Chapter 4, given delay tolerance level, switch-on and switch off decisions can be made by intelligent scheme. Therefore, improvements achieved by delayed access scheme in randomly operated network highlights the minimum possible achievements in energy efficiency.

We model the wireless channel by a standard propagation model with path loss exponent $\alpha > 2$ and Rayleigh fading with unit mean. In the typical case, the received power at a receiver is $P_{\mathrm{tx}}\, g_i r^{-\alpha}$, where $P_{\mathrm{tx}}(i)$ is transmission power of SC i, r_i is the distance between UE and SBS i, and g_i is the small-scale fading gain following exponential distribution with mean $1/\zeta$. Interference power at a receiver is the sum of the received powers from all active base stations excluding the

SBS that UE is associated with. Then, signal to interference and noise ratio is

$$\mathrm{SINR} = \frac{P_{\mathrm{tx}} h r^{-\alpha}}{\sum_{i \in \Phi_{\mathrm{A}} \setminus \{b_{\mathrm{o}}\}} P_{\mathrm{tx}} g_i r^{-\alpha} + \sigma^2}, \tag{5.1}$$

where σ^2 is the noise power of the additive noise, Φ_{A} denotes Homogenous Poisson Point Process for the active cells, and h is the small-scale fading gain associated with the channel between the user of interest and its serving SBS. We assume all active SBSs have constant transmit power, and interference from idle cells is negligible.

5.2 Analysis of Coverage Probability

In this section, we analyze the coverage probability of downlink transmission with delayed access in small cell networks. Analysis of coverage probability is important because it is highly critical for operators to sustain, with high probability, SINR levels above a threshold value that is acceptable for signal quality. Coverage probability defined as;

$$P_{\mathrm{c}} = \mathbb{P}(\mathrm{SINR} > \gamma), \tag{5.2}$$

where γ is the threshold SINR. By conditioning on the distance between UE and small cell, P_{c} can be written as

$$P_{\mathrm{c}} = \int_0^\infty \mathbb{P}(\mathrm{SINR} > \gamma | r) f_{r,w}(r) \mathrm{d}r, \tag{5.3}$$

where $f_{r,w}(r)$ is the probability density function of the distance to nearest idle cell and w is the tolerable delay. Let t be the access time of user to the cell. Clearly we have $t \in [0, w]$.

5.2.1 Distance Distribution to the Nearest Available Cell

Distance to the nearest cell can be evaluated by three independent events: immediate access within threshold distance (IA), delayed access within threshold distance (DA), access outside threshold distance (AO). Threshold distance is the communication range within which UE is willing to initiate commucation wiht SBS. The distance to the nearest base station is r. Then, the tail distribution of r in 2-D Poisson process can be written as:

$$\mathbb{P}[R < r] = P[\text{ BS closer than } r\ ] = 1 - e^{-\pi\rho_c r^2}, \tag{5.4}$$

and pdf of r is

$$f_r(r) = 2\pi\rho_c r^2 e^{-\pi\rho_c r^2}. \tag{5.5}$$

In case of IA, number of idle cells Poisson distributed with mean $N_\mathrm{I} = \Pi_\mathrm{I}\rho_f\pi R_\mathrm{th}^2$ for given R_th. Conditioning on number of idle cells outside an inner circle with radius r within transmission range, tail distribution of distance to the nearest idle cell can be written as

$$\begin{aligned}\mathbb{P}(R > r|R_\mathrm{th}, \mathrm{IA}) =& \frac{\sum_{i=1}^{\infty}\left[1 - \frac{r^2}{R_\mathrm{th}^2}\right]^i}{1 - e^{-N_\mathrm{I}}}\frac{e^{-N_\mathrm{I}}N_\mathrm{I}^i}{i!} \\ =& \frac{1}{1 - e^{-N_\mathrm{I}}}\left(e^{-N_\mathrm{I}\frac{r^2}{R_\mathrm{th}^2}} - e^{-N_\mathrm{I}}\right).\end{aligned} \tag{5.6}$$

Distance distribution is independent of the density of small cells in case of DA event. Due to random distribution of sleep, active cells, and the memoryless property of exponential distribution, any cell within R may become available irrespective of length of waiting time. Then, distance distribution to the cell is

$$\mathbb{P}(R > r|\ R_\mathrm{th}, \mathrm{DA}) = 1 - \frac{r^2}{R_\mathrm{th}^2}. \tag{5.7}$$

In case of OA, CDF of distance to the nearest cell is

$$\mathbb{P}(R > r|R_\mathrm{th}, \mathrm{OA}) = \frac{e^{-\rho_c\pi(r^2 - R_\mathrm{th}^2)}}{e^{-N_\mathrm{I}}}. \tag{5.8}$$

After taking the derivative of (5.6)-(5.8) with respect to r, pdf of distance can be formed as:

$$f_{R_{\mathrm{th}}}(r) = \begin{cases} \frac{e^{-\Pi_{\mathrm{I}}\rho_{\mathrm{c}}\pi r^2}\Pi_{\mathrm{I}}\rho_{\mathrm{c}}2\pi r}{1-e^{\Pi_{\mathrm{I}}\rho_{\mathrm{c}}\pi R_{\mathrm{th}}^2}} & : \mathrm{IA}, (t = 0) \\ \frac{2r}{R_{\mathrm{th}}^2} & : \mathrm{DA}, (0 < t < w) \\ \frac{e^{-Pi_{\mathrm{I}}\rho_{\mathrm{c}}\pi r^2}\Pi_{\mathrm{I}}\rho_{\mathrm{c}}2\pi r}{e^{\Pi_{\mathrm{I}}\rho_{\mathrm{c}}\pi R_{\mathrm{th}}^2}} & : \mathrm{OA}, (t = w) \end{cases} \cdot \tag{5.9}$$

The probability distribution of distance to nearest cell is not only a function of idle cell density but also the waiting time. Taking into consideration of tolerable delay and by the UE's access time t to the small cell, the respective probabilities of IA, DA, AO events can be written as

$$\mathbb{P}(\mathrm{IA}) = 1 - e^{-\Pi_{\mathrm{I}}\rho_{\mathrm{c}}\pi R_{\mathrm{th}}^2} \tag{5.10}$$

$$\mathbb{P}(\mathrm{DA}) = e^{-\Pi_{\mathrm{I}}\rho_{\mathrm{c}}\pi R_{\mathrm{th}}^2} - e^{-\beta_{\mathrm{w}}\rho_{\mathrm{c}}\pi R_{\mathrm{th}}^2} \tag{5.11}$$

$$\mathbb{P}(\mathrm{OA}) = e^{-\beta_{\mathrm{w}}\rho_{\mathrm{c}}\pi R_{\mathrm{th}}^2}, \tag{5.12}$$

where β_w is the probability that a cell is either available or will become available within tolerable delay time w (i.e., $\beta_0 = \Pi_{\mathrm{I}}$), which can be easily derived as

$$\beta_w = 1 - \Pi_{\mathrm{A}}e^{-\mu w_{\mathrm{t}}} - \Pi_{\mathrm{S}}e^{-\lambda_{\mathrm{S}} w_{\mathrm{t}}}. \tag{5.13}$$

5.2.2 Distribution of Coverage Probability

Coverage probability can be found by conditioning on distance. Using piece-wise density functions in (5.9), and (5.10)-(5.12), we can re-write (5.3) as

$$\begin{aligned}
\mathbb{P}(\mathrm{SINR} > \gamma) = & \\
& \int_0^{R_{\mathrm{th}}} \mathbb{P}(\mathrm{SINR} > \gamma | r) f_{R_{\mathrm{th}}|\mathrm{IA}}(r,t)\mathrm{d}r \times \mathbb{P}\left(\mathrm{IA}\right) \\
& + \int_0^{R_{\mathrm{th}}} \mathbb{P}(\mathrm{SINR} > \gamma | r) f_{R_{\mathrm{th}}|\mathrm{DA}}(r,t)\mathrm{d}r \times \mathbb{P}\left(\mathrm{DA}\right) \\
& + \int_{R_{\mathrm{th}}}^{\infty} \mathbb{P}(\mathrm{SINR} > \gamma | r) f_{R_{\mathrm{th}}|\mathrm{OA}}(r,t)\mathrm{d}r \times \mathbb{P}\left(\mathrm{OA}\right), \qquad (5.14)
\end{aligned}$$

The coverage probability conditioned on the distance as in (5.3) can then be derived as follows:

$$\begin{aligned}
\mathbb{P}(\mathrm{SINR} > \gamma | r) &= \mathbb{P}\left\{\frac{P_{\mathrm{tx}} h r^{-\alpha}}{P_{\mathrm{tx}} I_{\Phi_{\mathrm{A}}} + \sigma^2} > \gamma\right\} \\
&= \mathbb{E}_{\mathrm{I}_{\Phi_{\mathrm{A}}}}\left[\mathbb{P}\left\{h > \frac{\gamma r^{\alpha}}{P_{\mathrm{tx}}}\left(P_{\mathrm{tx}} I_{\Phi_{\mathrm{A}}} + \sigma^2\right)\right\}\right] \\
&= \mathbb{E}_{\mathrm{I}_{\Phi_{\mathrm{A}}}}\left[\exp\left(-\frac{\zeta\gamma r^{\alpha}}{P_{\mathrm{tx}}}\left(P_{\mathrm{tx}} I_{\Phi_{\mathrm{A}}} + \sigma^2\right)\right)\right] \\
&= e^{-\frac{\zeta\gamma r^{\alpha}\sigma^2}{P_{\mathrm{tx}}}} \mathcal{L}_{\mathrm{I}_{\Phi_{\mathrm{A}}}}(\zeta\gamma r^{\alpha}), \qquad (5.15)
\end{aligned}$$

where $\mathcal{L}_{\mathrm{I}_{\Phi_{\mathrm{A}}}}(s)$ is the Laplace transform of $I_{\Phi_{\mathrm{A}}}$ conditioned upon the transmit distance r. For $\mathcal{L}_{\mathrm{I}_{\Phi_{\mathrm{A}}}}(s)$, we have

$$\begin{aligned}
\mathcal{L}_{\mathrm{I}_{\Phi_{\mathrm{A}}}}(s) &= \mathbb{E}_{\Phi_{\mathrm{A}}}\left[e^{-s\sum_{i\in\Phi_{\mathrm{A}}} g_i r_i^{-\alpha}}\right] \\
&= \mathbb{E}_{\Phi_{\mathrm{A}}}\left[\prod_i e^{-s g_i r_i^{-\alpha}}\right] \\
&= \exp(-\Pi_{\mathrm{A}} 2\pi\rho_{\mathrm{c}} \int_0^{\infty} \left(1 - \mathbb{E}_{g_i}\left[-s g_i z_i^{-\alpha}\right]\right)) z \mathrm{d}z \\
&= \exp\left(-\Pi_{\mathrm{A}}\pi\rho_{\mathrm{c}} r^2 \gamma^{\frac{2}{\alpha}} \int_0^{\infty} \frac{\mathrm{d}z}{1 + z^{\frac{\alpha}{2}}}\right). \qquad (5.16)
\end{aligned}$$

Third equality is due to $\mathbb{E}[\prod_{x\in\Phi} f(x)] = e^{-\int_{\mathbb{R}^2}(1-f(x)\rho \mathrm{d}x}$ [113]. In the last equality of (5.16), we plug $s = \zeta\gamma r^{\alpha}$. As a special case for $\alpha = 4$, one can easily find,

$$\mathcal{L}_{\mathrm{I}_{\Phi_{\mathrm{A}}}}(\zeta\gamma r^4) = e^{-\Pi_{\mathrm{A}}\rho_{\mathrm{c}} r^2\sqrt{\gamma}\pi^2/2}. \tag{5.17}$$

We can further derive closed-form solutions with and without white noise.

5.2.3 Special Cases

Theorem 2. *For $\alpha = 4, \sigma^2 = 0$, coverage probability is*

$$\begin{aligned} P_{\mathrm{c}} = {} & \frac{\beta_0}{\eta+\beta_0} \\ & + \left(e^{-\beta_0\nu_c} - e^{-\beta_{\mathrm{w}}\nu_c}\right)\left[\frac{1}{\eta\nu_c} - e^{-\eta\nu_c}\left\{\frac{\beta_0}{\eta+\beta_0} + \frac{1}{\eta\nu_c}\right\}\right], \end{aligned} \tag{5.18}$$

Proof. It can be easily seen that (6.13) and (5.15) becomes equal when $\sigma^2 = 0$. Then, by inserting (5.10)-(5.12), and (6.13) into (5.14), we have:

$$\begin{aligned} P_c = {} & \beta_0 \int_0^{\nu_c} e^{-u(\beta_0+\eta)}\mathrm{d}u \\ & + \frac{1}{\nu_c}\int_0^{\nu_c} e^{-\eta u}\mathrm{d}u \left(e^{-\beta_0\nu_c} - e^{-\beta_w\nu_c}\right) \\ & + \frac{\beta_0 e^{-\beta_w\nu_c}}{e^{-\beta_0\nu_c}}\int_{\nu_c}^{\infty} e^{-(\eta+\beta_0)u}\mathrm{d}u, \end{aligned} \tag{5.19}$$

where $\eta = \Pi_{\mathrm{A}}\sqrt{\gamma}\pi/2$, and $\nu_c = \rho_{\mathrm{c}}\pi R_{\mathrm{th}}^2$. After some algebraic manipulations, coverage probability becomes

$$\begin{aligned} P_{\mathrm{c}} = {} & \frac{\beta_0}{\eta+\beta_0} \\ & + \left(e^{-\beta_0\nu_c} - e^{-\beta_{\mathrm{w}}\nu_c}\right)\left[\frac{1}{\eta\nu_c} - e^{-\eta\nu_c}\left\{\frac{\beta_0}{\eta+\beta_0} + \frac{1}{\eta\nu_c}\right\}\right]. \end{aligned} \tag{5.20}$$

□

Similarly, interference with noise ($\alpha = 4$) can be derived as in 5.20. Plugging (6.13) in (5.15), then, substituting (5.10)-(5.12), (5.15) into (5.14), coverage probability is derived as:

$$\begin{aligned} P_{\mathrm{c}} = \frac{1}{\sigma}\sqrt{\frac{\pi P_{\mathrm{tx}}}{\zeta\gamma}}\Big(\Pi_{\mathrm{I}}\pi\rho_{\mathrm{c}}K_1\left[\frac{1}{2} - K_3(1 - e^{-(\beta_w-\beta_0)\nu_c})\right] \\ + \frac{1}{R^2}K_2(K_4 - \frac{1}{2})(e^{-\beta_0\nu_c} - e^{-\beta_W\nu_c})\Big), \end{aligned} \tag{5.21}$$

where K_{1-4} are given by:

$$K_1 = e^{\frac{(\Pi_{\mathrm{A}}\sqrt{\gamma}\pi/2-\Pi_{\mathrm{I}})^2(\rho_{\mathrm{c}}\pi)^2 P_{tx}}{4\zeta\gamma\sigma^2}}, \tag{5.22}$$

$$K_2 = e^{\frac{(\Pi_A\sqrt{\gamma}\pi^2)^2 P_{\mathrm{tx}}}{16\zeta\gamma\sigma^2}}, \tag{5.23}$$

$$K_3 = Q\Big(\frac{2\sigma}{\rho_{\mathrm{c}}\pi}\sqrt{\frac{\zeta\gamma}{P_{\mathrm{tx}}}}(\nu_c + \frac{(\Pi_{\mathrm{A}}\gamma\pi/2 + \Pi_{\mathrm{I}})(\rho_{\mathrm{c}}\pi)^2 P_{\mathrm{tx}}}{4\zeta\gamma\sigma^2})\Big), \tag{5.24}$$

$$K_4 = \Phi\left(\frac{2\zeta\gamma\sigma^2}{P_{\mathrm{tx}}}\left(R^2 + \frac{p_{rmA}\rho_{\mathrm{c}}\gamma\pi^2 R_{\mathrm{th}}^2 P_{\mathrm{tx}}}{4\zeta\gamma\sigma^2}\right)\right). \tag{5.25}$$

5.3 Average Bit Rate for Delayed Access

In this section, we derive the average achievable bit-rate as a function of threshold distance and waiting time by using the respective probabilities of IA, DA and OA (5.10)-(5.12). For capacity, we used Shannon's capacity bound, which is widely used in many studies. To be more precise, we derive expected value of $\log_2$(1+SINR), which gives us achievable capacity in terms of bits/hz. Conditioning on access types depending on waiting time of DE, we have

$$\begin{aligned} \tau = \mathbb{E}\left[\log_2(1 + \mathrm{SINR})\right] = \mathbb{E}\left[\log_2(1 + \mathrm{SINR}) \,|\, \mathrm{IA}\right]\mathbb{P}(\mathrm{IA}) \\ + \mathbb{E}\left[\log_2(1 + \mathrm{SINR}) \,|\, \mathrm{DA}\right]\mathbb{P}(\mathrm{DA}) \\ + \mathbb{E}\left[\log_2(1 + \mathrm{SINR}) \,|\, \mathrm{OA}\right]\mathbb{P}(\mathrm{OA}). \end{aligned} \tag{5.26}$$

For each access type (i.e. IA, DA, or OA),conditional expectation needs to be derived. Here, we start with average bitrate if UE connects to the cell immediately:

$$\mathbb{E}\left[\log_2(1{+}\mathrm{SINR})\,|\,\mathrm{IA}\right] = \\ \int_0^{R_{\mathrm{th}}} f_{\mathrm{IA}}(r,t) \int_0^{\infty} \frac{\ln(1{+}\gamma)}{\ln(2)} f_{\gamma|r}(\gamma)\mathrm{d}\gamma\mathrm{d}r, \tag{5.27}$$

where f_{IA} is the density function of distance when DE connects immediately, and $f_{\gamma|r}(\gamma)$ is the conditional density function of SINR given that distance betwen user and transmitting cell is r. In the inner integration in (6.14), we can write the conditional pdf using the tail distribution of SINR in (5.15). Then, we have:

$$\begin{aligned}
\tau &= \frac{1}{\ln(2)} \int_0^{R_{\mathrm{th}}} f_{\mathrm{IA}}(r,t) \int_0^{\infty} \ln(1{+}\gamma)\mathrm{d}\left(\mathbb{P}\left(\mathrm{SINR} > \gamma|r\right)\right)\mathrm{d}r \\
&= \frac{1}{\ln(2)} \int_0^{R_{\mathrm{th}}} f_{\mathrm{IA}}(r,t) \int_0^{\infty} \frac{\mathbb{P}\left(\mathrm{SINR} > \gamma|r\right)}{1+\gamma}\mathrm{d}\gamma\mathrm{d}r \qquad (5.28) \\
&= \frac{1}{\ln(2)} \int_0^{\infty} \left(\int_0^{R_{\mathrm{th}}} \frac{\mathbb{P}\left(\mathrm{SINR} > \gamma|r\right)}{1+\gamma} f_{\mathrm{IA}}(r,t)\mathrm{d}r \right)\mathrm{d}\gamma. \qquad (5.29)
\end{aligned}$$

The second equality of (6.15) is due to change of variables in the inner integration. After changing the order of integration in third equality, the inner integration becomes equivalent to (5.14). Following similar procedure for $\mathbb{E}\left[\log_2(1 + \mathrm{SINR}\ )\,|\,\mathrm{DA}\right]$, and $\mathbb{E}\left[\log_2(1 + \mathrm{SINR}\ )\,|\,\mathrm{OA}\right]$, and plugging them in (5.26), we can obtain the capacity as:

$$\tau = \frac{1}{\ln(2)} \int_0^{\infty} \frac{\mathbb{P}\left(\mathrm{SINR} > \gamma\right)}{1+\gamma}\mathrm{d}\gamma\ . \tag{5.30}$$

5.4 Optimization of Bit Rate and Coverage Probability for Delayed Access

In this section, we optimize the coverage probability and average bit rate with respect to threshold distance. Before delving into analytical details, it is helpful to explain why an optimal threshold distance exists. Consider two extreme cases of small and large threshold distances. In the first case, it is not likely that UE finds an idle cell within its threshold distance, causing waste of delay budget. For a large threshold distance, a user is very likely to access an idle cell without much delay, causing minimal use of delay budget. None of these choices yields optimal coverage probability with respect to waiting time. Therefore it is necessary to adjust threshold distance with respect to given delay budget.

5.4.1 Optimization of Coverage Probability

In this section, we optimize coverage probability with respect to a given tolerable delay by adjusting the threshold distance. For analytical tractability, we made several assumptions. First, we assumed that noise power is negligible when compared with interference power. Second, we assumed cells can adjust their coverage range by increasing or decreasing their transmit power. Notice that coverage probability in (5.20) is not a function of transmit power. In other words, UE's access strategy does not affect interference power.

For the analysis, we only considered a special case of $\alpha = 4$. To find threshold distance R_{th} that maximizes coverage probability, we first take the derivative of (5.20) respect to ν_c. We then obtain:

$$\begin{aligned}\frac{\mathrm{d}P_{\mathrm{c}}}{\mathrm{d}\nu_c} &= \beta_0 e^{-(\beta_0+\eta)\nu_c} \\ &- \frac{1}{\eta{\nu_c}^2}\Big(e^{-\beta_0\nu_c} - e^{-\beta_w\nu_c} - e^{-(\beta_0+\eta)\nu_c} + e^{-(\beta_w+\eta)\nu_c}\Big) \\ &+ \frac{1}{\eta\nu_c}\Big(- \beta_0 e^{-\beta_0\nu_c} + \beta_w e^{-\beta_w\nu_c} + (\beta_0+\eta)e^{-(\beta_0+\eta)\nu_c} \\ &- (\beta_w+\eta)e^{-(\beta_w+\eta)\nu_c}\Big) + \frac{\beta_0(\beta_w+\eta)}{\beta_0+\eta}e^{-(\beta_w+\eta)\nu_c}. \end{aligned} \tag{5.31}$$

After substituting second order Taylor expansion for $e^{-\beta_0\nu_c}$, $e^{-\beta_w\nu_c}$, $e^{-(\beta_0+\eta)\nu_c}$, $e^{-(\beta_w+\eta)\nu_c}$, derivation in (5.31) reduces to quadratic form of $m_1{v_c}^2 + m_2 v_c + m_3$, where m_1, m_2, m_3 are

$$\begin{aligned} m_1 &= \beta_0 \left(\frac{(\beta_0 + \eta)^2}{2} + \frac{(\beta_w + \eta)^3}{2(\beta_0 + \eta)} \right) \\ m_2 &= \frac{3}{2} (\beta_0 - \beta_w)(\beta_0 + \beta_w + \eta)) - \beta_0 \left(\beta_0 + \eta + \frac{(\beta_w + \eta)^2}{\beta_0 + \eta} \right) \\ m_3 &= 2(\beta_w - \beta_0) + \beta_0 \left(1 + \frac{\beta_w + \eta}{\beta_0 + \eta} \right). \end{aligned} \tag{5.32}$$

Let J_1^*,J_2^* be the roots of the quadratic quadratic equation in (5.32). Since $\beta_\mathrm{w} > \beta_0$, we have $m_1 > 0$, $m_2 < 0$, $m_3 > 0$, and thus J_1^*,J_2^* are both positive. Then, candidate threshold distances $R^*_{\mathrm{th}_1}$, $R^*_{\mathrm{th}_2}$ are $\sqrt{\frac{J_1^*}{\pi\rho_\mathrm{c}}}$, $\sqrt{\frac{J_2^*}{\pi\rho_\mathrm{c}}}$. Finally, among the two threshold distances, we choose the one satisfying second derivative test. To satisfy maximum condition, the second derivative of (5.32) necessitates $v_c < \frac{m_2}{2m_1}$, which is met by the smaller root.

Lemma 1. *Optimal threshold distance that maximizes the coverage probability is a decreasing function of* γ.

Proof. We have $\gamma \sim \eta^2$. From (5.32), it is clear that $m_1 \sim \Theta(\eta^3)$, $m_2 \sim \Theta(\eta^2)$, $m_3 \sim c$, where c is a constant. Then, the roots $\frac{-m_2 \pm \sqrt{m_2^2 - 4m_1m_3}}{2m_1}$ are decreasing with respect to η.

5.4.2 Optimization of Bit-rate

Optimizing the bitrate with respect to threshold distance is not easy to derive analytically. Therefore, we developed an efficient numerical algorithm that computes the optimal threshold distance according to the bitrate.

Algorithm initializes with an upper and a lower bound of threshold distance R_u, R_l respectively. Simply, we choose $R_\mathrm{l} = 0$. Derivation of upper bound can be done based on the threshold distance maximizing the coverage probability. We observe from Eq. (5.30) that maximizing the coverage probability for given SINR value is strongly related to maximizing bitrate. Then, without loss of generality we can rewrite Eq. (5.30) in a discrete form as $\lim_{\Delta\gamma\to 0} \sum_{i=1}^{\infty} \frac{\mathbb{P}(\mathrm{SINR}>\gamma_i)}{1+\gamma_i} \Delta\gamma_i$. Let R_i^* be the optimal threshold distance for $\mathbb{P}(\mathrm{SINR} > \gamma_i)$. Then, by choosing optimal thresholds R_i^*, sum of the bit rates corresponding to each γ_i is maximized. By lemma 5.4.1, we have $R_i^* > R_j^*$, where $i < j$. Finally by choosing arbitrarily small γ value, an upper bound can be defined. After we define R_l, R_u, we can easily find the optimal threshold distance by following Algorithm 5.

Algorithm 5 Algorithm to find optimal R_th for bitrate.

```
Initialize parameters $\varepsilon$, $R_\mathrm{l} = 0$, $R_\mathrm{u}$, $\Delta R = R_u$
while $\Delta R > \varepsilon$ do
    Set $R = (R_l + R_u)/2$ and find search direction by computing bitrates $r_{R+\varepsilon}$, $r_{R-\varepsilon}$
    if $r_{R+\varepsilon} > r_{R-\varepsilon}$ then
        $R_\mathrm{l} \leftarrow R$
    else
        $R_\mathrm{u} \leftarrow R$
    end if
    $\Delta R \leftarrow |R - (R_u + R_l)/2|$
end while
```

Note that Algorithm 1 has a precision parameter ε. At each iteration, candidate threshold distance is chosen as the average value of upper and lower bounds of R_th. Also, search direction is determined by checking the bit-rates in $\varepsilon-$neighborhood of R. The algorithm eventually stops and converges after $\Theta(\log(\frac{R_u}{\varepsilon}))$ steps.

5.5 Energy Efficiency Analysis

In this part, we consider the trade-off between energy-efficiency and delay. Energy-efficiency is measured as the amount of transmitted bits per unit time per unit bandwidth per Watt (i.e. bits/s/Hz/Watt). Then energy-efficiency is

$$EE(\rho_\mathrm{c}, \beta_w) = \frac{\tau(\rho_\mathrm{c}, \beta_\mathrm{w})}{\mathrm{B} \times \text{Mean power consumption of a cell}},$$

where B is the bandwidth. Without loss of generality, we assume SBS uses all available bandwidth during transmission. In order to compare energy efficiency of a cell with and without delay, we define the normalized energy efficiency as

$$EE_N = \frac{EE(\rho_c, \beta_w)}{EE(\rho_c, \beta_0)}, \tag{5.33}$$

which measures relative scale of improvement in the energy-efficiency with respect to the condition with no access delay (i.e., $t = 0$).

5.6 Simulation and Discussions

In this part, we verify the accuracy of our analysis by Monte-Carlo simulations. We generate uniformly distributed random variables for x- and y- coordinates of small cells in a large circle. Cells are independently marked as either idle, active or sleeping with probabilities Π_I, Π_A and Π_S, respectively. Simulation parameters are listed in Table 5.1.

Table 5.1: Simulation Parameters.

Parameter	Value
Small Cell density (ρ_c)	0.005 per m^2
Tolerable delay (w_t)	10 sec^{-1}
Small Cell Transmission Power (P_{tx})	23 dBm
Thermal Noise Power (σ^2)	-104 dBm
Path loss exponent (α)	4
Average sleep/active time (λ_S)	10 sec
Average sleep/active time (μ)	10 sec

In Figure 5.1, coverage probability derived in Eq. (5.20) is verified via simulations. We observe that choosing optimal threshold distance with respect to SINR threshold has significant importance. Especially at low SINR levels, the selection of threshold distance becomes crucial. We observe that coverage probability at 0 dB SINR is almost doubled by choosing proper threshold distance. Especially for delay tolerant data applications, and for intermittent connections, delayed access mechanisms can sustain high coverage probability and improves energy efficiency.

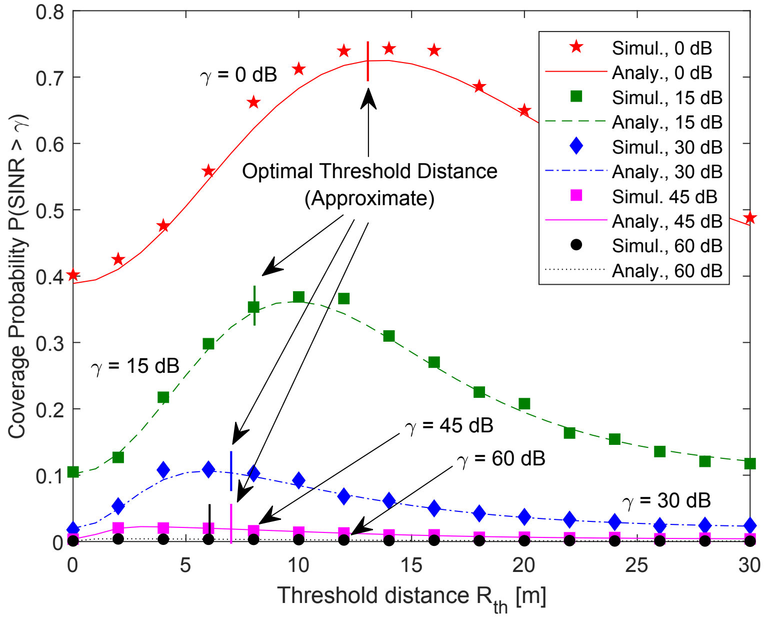


Figure 5.1: Coverage probability at different SINR thresholds with no noise. Threshold distance with respect to given threshold SINR is close to optimum. The small gap between approximation and optimal value is due to Taylor approximation. Tolerable delay $w_{\mathrm{t}} = 1/\lambda_{\mathrm{S}}$.

Figure 5.2 shows capacity increase with respect to threshold distance. Tolerable delay is normalized by the sleeping time. Depending on the fraction of sleep and idle cells, and the improvement in capacity is almost tripled. Capacity gain depends on the transmit power gain. In fact, it is shown in Chapter 3 that transmit power gain is a function of the idle cell density. If the idle cell density is small with respect to active and sleep cell densities, potential transmit power gain is large. In theory, idle cell density can be arbitrarily small compared to active and sleep cell density. We also observe that optimal ranges with respect to different delay values are close. However, as UE's tolerable delay increases, optimal range also decreases because UE is able to find a closeby SBS by waiting longer. SBS density is for dense femtocell deployment so average distance to SBS is $\frac{1}{\sqrt{\rho_{\mathrm{c}}}} \approx 14\ \ \mathrm{m}$. If we decrease the density, we expect tolerable delay to play a more significant role in maximizing throughput.

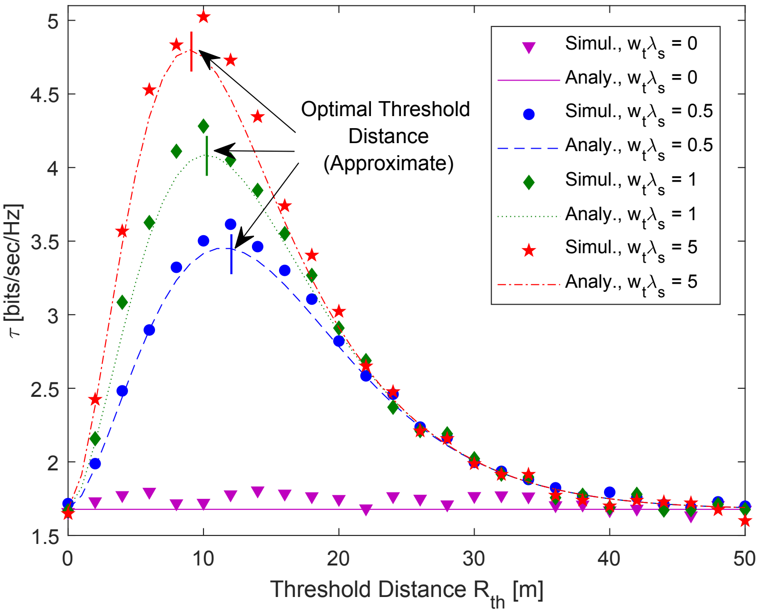


Figure 5.2: Transmission rate for varying threshold distances and delay. Small vertical lines show optimal threshold distance computed by our algorithm.

Figure 5.3 shows the improvements of the network energy efficiency with respect to delay. Specifically, simulation starts with delay intolerant condition, i.e., $\beta_w = \beta_0 = 0.1$. As the tolerable delay increases, the proportion of cells that become available, either in time or immediately, β_w, the energy-efficiency of network improves. At high or moderate traffic it is detrimental to turn off cells due to the possibility of degrading quality of service. Because of this, we evaluated network energy efficiency only at low traffic utilization (i.e., $\Pi_A = 0.1$). For simulation, we kept density of active cells small and fixed and changed the density of sleeping and idle cells. We observe that as the density of sleeping cells with respect to the density of idle cells increases, energy-efficiency of the small cell network also increases significantly.

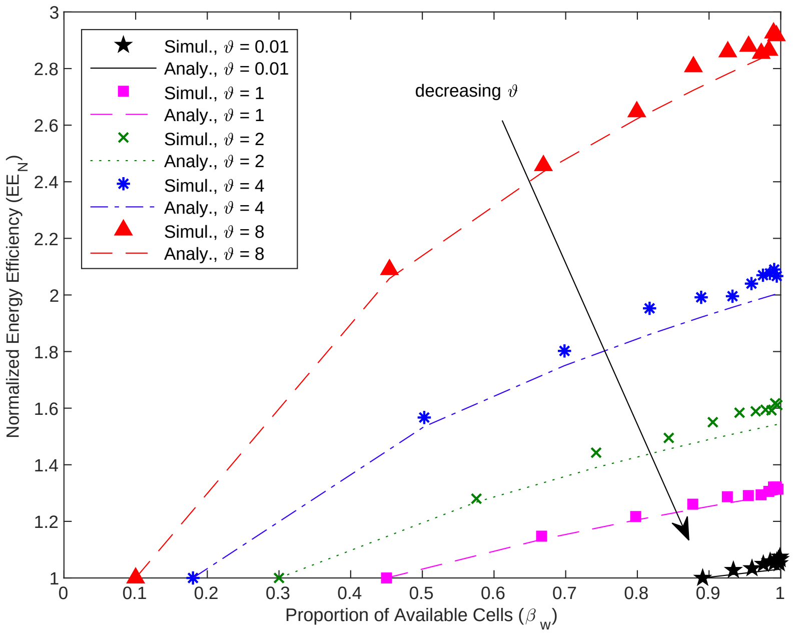


Figure 5.3: Normalized energy efficiency with respect to cell availability along with $\vartheta = \Pi_S/\Pi_I \in \{0.01, 1, 2, 4, 8\}$ and $\Pi_A = 0.1$.

5.7 Conclusion

In this chapter, delayed access mechanism is analyzed in large scale network considering the metrics such as energy efficiency, coverage probability, and average bit-rate. Analysis is verified via simulations. Optimal threshold distance maximizing the coverage probability is derived. An efficient numerical algorithm that optimizes threshold distance with respect to the delay budget is developed. Energy-efficiency of the small cell network is assessed by comparing UE's connection to the cell with and without delay. Results show that delayed access strategy can be utilized for data applications that can tolerate an initial access delay. By delaying UE's access, energy-efficiency of the network can be increased significantly at low traffic utilization. Some of our future work include designing on/off schemes with deterministic sleep times, and development of energy efficient protocols for small cell networks.

Chapter 6: Delay Spectral Efficiency Tradeoff in Emerging Applications Using Edge Service

Some CPU hungry applications hold the potential to dominate wireless traffic; thefore, their effect on small cell networks needs to be investigated. According to the forecast reports, number of virtual reality (VR) and augmented reality (AR) devices sold will reach about 99 million in 2021 [114]. VR/AR industry will reach $108 billion [115]. Virtual reality and augmented reality applications require high computation sources which are envisioned to be met by edge computing services. What is more, available off-the shelf (i.e. HTC Vive, Oculus Rift) may operate at 2160x1200 resolution and 90 fps. Amount of traffic load may require up to 5.2 Gbps, which is far above the supported speeds in the current wireless infrastructure. [116]. Moreover, offshore oilrigs generate about 500 GB data in a week. Commercial jet airplanes generate 10 TB data in a 30 minute period during flight [117]. Shay states [118] due to bandwidth limitations only 20% of data can be processed at cloud platform. So, there is a risk that wireless network may be overloaded because of huge amounts of data transfer to remote processing units.

Dynamics in wireless medium such as noise, physical obstacles, traffic fluctuations challenge the seamless operation of VR applications. At high traffic load period or during bad channel conditions, tight delay constraints of these applications may not be met. For example, although milimeter wave channels can achieve Gbs speeds, their drastic attenuation due to temporary and stationary blockages is a big concern to realize remote rendering for VR applications. Therefore, local processing is necessary to guarantee quality of service need of computer intensive applications. Moreover, due to requiring high speed, network operators may limit VR applications during high traffic periods for the sake of fairness and avoidance of congestion, favoring the usage of local processing.

There are many studies regarding edge/fog processing. In [119], an energy optimal task offload schme is introduced. Gilbert-Elliott (GE) model (i.e., two state markov chain for low and high gains), is considered for the wireless link. In [116], an experiment set-up that remotely renders of VR frames is tested. Results are evaluated for line of sight scnenario in milimeter wave channel. In [120], VR system is assumed to be capable of connecting multiple access points equipped with antenna using milimeter wave frequencies. Frame caching and multi point transmission is offered to resolve the latency in case of high attenuation in milimeter wave links. In [121], a solution to remote the processing problem with reliabilty constraints is offered. It is assumed that UE can offload to multiple edge processors simultaneously, and burden on wireless links can be lifted by *cacheable* contents such as object detection results, which can be used at a later time.

In Chapters 3, and 4, our main objective was to make small cell networks energy effiecent. Energy conservation is achieved by turning cells off and utilizing user equipment's access delay. In Chapter 5, delay capacity tradeoff is analyzed. The scope of this chapter aims to give a modest analysis about the impact of bandwidth hungry and CPU intensive applications on small cell networks. Specifically, we consider computer intensive tasks that are either processed local CPU or at edge server. If tasks are processed locally, there will be delay due to limited processing capability of user equipment. There will be various delays in case of remote processing such as transmission delay, queuing delay at cloudlet (i.e., mechanisms for authentication, access control, pricing, resource allocation), and processing delay at edge/fog processor. There is fundamental tradeoff between delay and bandwidth due to increasing traffic volume of applications that consume large bandwidth, and require high processing service.

Section 6.1 discusses details of our system model. In Section 6.2, performance evaluation of local processing and spectral efficiency of small cell network are analyzed. Finally, simulation results are discussed in Section 6.3.

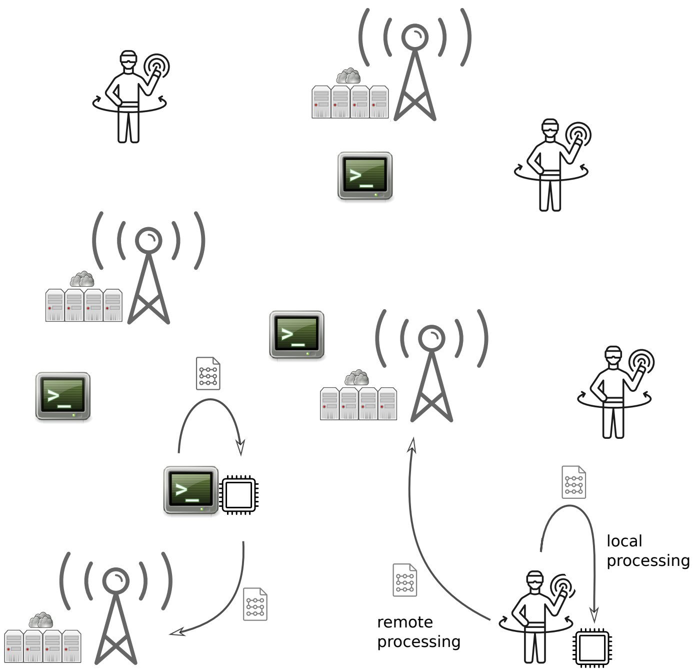


Figure 6.1: Operation of local computing at the UE versus edge computing through small cell network. CPU-intensive tasks can be computed either locally at the UE or remotely at the edge/fog depending on the tolerable delay of the tasks.

6.1 System Model

In this section, we introduce the system model. UEs have computationally intensive tasks such as VR, AR frames which can be processed either locally or remotely at edge server (see Figure 6.1). Tasks arrival follows a poisson process with rate λ_{u}. Distribution of UEs and SBSs are HPPP as in previous prior chapters. Each task can tolerate a delay with rate λ_{W}. Task arrivals join to the internal queue, and wait to be processed locally. Only tasks that are not processed within a tolerable delay are transmitted to the edge server.

If the task is not processed locally and tolerable delay expired, UE immediately offloads the task to edge server by accessing the nearest idle cell in the small cell network. We do not employ delayed access strategy because it will introduce additional latency. Besides, all small base stations are on. The rationale behing keeping all SBS fully powered is to meet strict performance and

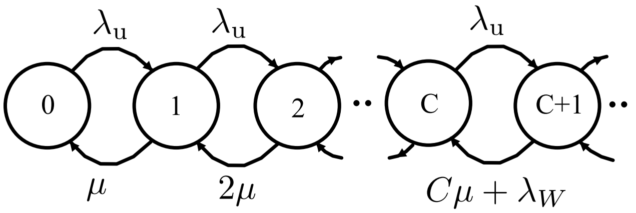


Figure 6.2: Operation of UEs that process delay-sensitive tasks. Tasks whose tolerable delay expires leave the queue.

tight delay requirements of VR applications. Utilization of network will be medium or high, and therefore not suitable to apply energy saving schemes.

Service time of task has mean $1/\mu_{\mathrm{l}}$ if processed locally. The local CPU is modelled as $M/M/c$ queue with impatience as shown in Figure 6.2. Total end-to-end delay is assumed following exponential distribution with mean $1/\mu_{\mathrm{f}}$. Note that end-to-end delay is comprised of various delay components such as buffering delay at uplink scheduler, task upload period, buffering delay at edge server, processing period at edge server, downlink buffering delay, and downlink transmission period. Distributions of these delays depend on utilization of wireless network, edge computing resources and type of task that requires remote processing. According to [120], acceptable latency for typical VR application is about 25 ms.

It requires a very complex analysis to model end-to-end delay due to various latency components. Therefore, we focus our model to specific problem or virtual reality. Considering typical VR application, UE offloads location coordinates to the edge, and VR frame is generated at edge processor. Therefore, the size of uploaded data is small with respect to frame size to be downloaded. It is even possible to predict head movements [122] so that sum of the delays at uplink scheduler and task upload period can be considered small. When the computing resources are sufficient, the prominent delay component is the downlink transmission. We expect that $1/\mu_{\mathrm{f}} \leq 1/\mu_{\mathrm{l}}$ when task download time is reasonably small and edge computing resources are sufficient. When the channel conditions are bad, or network resources are diminished, the end-to-end response time is longer than local processing time.

6.2 Analysis of M/M/c Queue and Sectral Efficiency of Small Cell Network

In this section, details of the analysis of the model is given. First, the queuing model in Figure 6.2 is analyzed. The aim of this analysis is simply to find the proportion of tasks that were completed within tolerable delay time. Tasks that are not completed in tolerable delay time are sent to the edge server through the small cell network. We analyze spectral efficiency of small cell network due to traffic load of uncompleted tasks.

6.2.1 Analyisis of M/M/c Queue With Exponential Impatience

Using equilibrium conditions, state probabilities of the queue shown in Figure 6.2 can be derived as follows:

$$p_n = \begin{cases} \frac{\lambda_{\mathrm{u}}^n}{n!\mu^n} p_0 & n < C, \\ \\ \frac{\lambda_{\mathrm{u}}^n}{\mu^{C-1}(C-1)\prod_{k=0}^{n-C}(c\mu+k\lambda_{\mathrm{W}})} p_0 & n \geq C, \end{cases} \tag{6.1}$$

where

$$p_0 = \left[\sum_{i=0}^{C-1} \frac{\lambda_{\mathrm{u}}^n}{n!\mu^n} + \sum_{n=C}^{\infty} \frac{\lambda_{\mathrm{u}}^n}{\mu^{C-1}(C-1)\prod_{k=0}^{n-C}(c\mu + k\lambda_{\mathrm{W}})} \right]^{-1}. \tag{6.2}$$

6.2.1.1 Proportion of tasks computed locally

It is necessary to find the fraction of tasks computed locally and the remaining tasks computed remotely. To find that, we further analyze the queue in Figure 6.2. Let q_n be the conditional probability that the arriving task “sees" the queue in state n, and successfully receives service within its tolerable delay. It is obvious that $q_i = 1$ for $0 \leq i \leq C - 1$. For $n = C$, we have $q_n = \frac{C\mu}{C\mu+\lambda_{\mathrm{W}}}$. In more general case (i.e., $C \geq n$), we want $q_n = P\{S_n < W|\, n\}$, where S_n is the sum of n exponential random variables with $C\mu$, $C\mu+\lambda_w$, $C\mu+2\lambda_w$,...,$C\mu+(n-1)\lambda_w$ rates. Distribution of the sum of n exponentials is given by [123]:

$$f_{S_n} = f_{X_1+X_2+..+X_n} = \left(\prod_{i=1}^{n} \theta_i\right) \sum_{j=1}^{n} \frac{e^{-\theta_j t}}{\prod_{j\neq s;s=1}^{n}(\theta_s - \theta_j)}, \tag{6.3}$$

where $\Theta_n = [C\mu, C\mu + \lambda_w, .., C\mu + (n-1)\lambda_w]$. Then,

$$\begin{aligned} q_n &= \int_0^\infty \left(\int_0^w f_{S_n}(x)dx\right) \lambda_w e^{-\lambda_w w} dw \\ &= \left(\prod_{i=1}^{n} \theta_i\right) \sum_{j=1}^{n} \frac{\int_0^\infty \left(\int_0^w e^{-\theta_j x}dx\right) \lambda_w e^{-\lambda_w w} dw}{\prod_{j\neq s;s=1}^{n}(\theta_s - \theta_j)} \\ &= \left(\prod_{i=1}^{n} \theta_i\right) \sum_{j=1}^{n} \frac{1}{(\lambda_w + \theta_j)\prod_{j\neq s;s=1}^{n}(\theta_s - \theta_j)} \end{aligned} \tag{6.4}$$

and finally, considering each arrival we have:

$$P_{\mathrm{R}} = \sum_{n=0}^{\infty} P\{S_n < W|\, n\} p_n \tag{6.5}$$

6.2.1.2 Delay Distribution

In this part, we derive the distribution of the delay experienced at the queue. The distribution of delay can be found by conditioning on whether the UE receives service. Let R be the event that UE receives service. Then, we may write:

$$P(T < t) = P(T < t|R)(1 - P_B) + P(T < t|R^c)P_B, \tag{6.6}$$

which after some algebratic manipulation can be rewritten as:

$$\begin{aligned} P(T < t) &= P(S < t|S < W)(1 - P_B) + P(W < t|W < S)P_B \\ &= \left(1 - \sum_{r=0}^{\infty} P(S_r > t|S_r < W)\gamma_r\right)(1 - P_B) \\ &+ \left(1 - \sum_{r=0}^{\infty} P(W > t|W < S_r)\gamma_r^c\right)P_B\,, \end{aligned} \tag{6.7}$$

where $\gamma_n = \frac{q_n p_n}{(1-P_B)}$ and $\gamma_n^c = \frac{(1-q_n)p_n}{P_B}$. On the other hand, $P(S_n > t|S_n < W)$ is the conditional tail probability that an arriving request finds n tasks in the system and receives service no earlier than t. Then, we have:

$$\begin{aligned} P(S_n > t|S_n < W) &= \frac{P(t < S_n < W)}{P(S_n < W|n)} \\ &= \frac{\int_t^{\infty}\left[\int_t^{w} f_{S_n}(x)dx\right]\lambda_w e^{-\lambda_w w}dw}{P(S_n < W|n)} \\ &= \frac{1}{q_n}\sum_{j=1}^{n}\frac{1}{\prod_{s=1,j\neq s}^{n}(\theta_s - \theta_j)}\frac{e^{-t(\theta_j+\lambda_w)}}{\theta_j + \lambda_w}\ . \end{aligned} \tag{6.8}$$

Similarly, we obtain $P(W > t|W < S_n)$ by conditioning on the sum of delays as follows:

$$\begin{aligned} P(W > t|W < S_n) &= \frac{P(t < W < S_n)}{P(S_n > W)} \\ &= \frac{\int_t^{\infty}\left(\int_t^{w}\lambda_w e^{-\lambda_w x}dx\right)f_{S_n}(w)dw}{1 - q_n} \\ &= \frac{1}{1 - q_n}\left[\prod_{j=1}^{n}\theta_j\right]\times\sum_{j=1}^{n}\frac{\lambda_w e^{-(\theta_j+\lambda_w)t}}{\theta_j(\theta_j + \lambda_w)\prod_{k\neq j,k=1}^{(\theta_k-\theta_j)}}\ . \end{aligned} \tag{6.9}$$

6.2.1.3 Average Queuing Delay

Average delay, $E[D]$ due to queuing delay can be derived by conditioning on whether the UE receives service locally or remotely, as follows:

$$\begin{aligned} E[D] &= E[S|S < W](1 - P_B) + E[W|S > W]P_B \\ &= \left(\sum_{i=1}^{\infty} E[S_n|S_n < W]\gamma_n\right)(1 - P_B) + \left(\sum_{i=1}^{\infty} E[W|W < S_n]\gamma_n^c\right)P_B \ . \end{aligned}$$

We obtain the conditional expectation $E[S|S_n < W]$, and $E[W|W < S_n]$ by integrating (6.8) and (6.9), respectively, with respect to t.

6.2.2 Spectral Efficiency Analysis

In this section, we analyze how spectral efficiency of small cell network changes with respect to the delay tolerance of the application. Measure of spectral efficiency is b/s/Hz. First, we give an intuitive explanation how spectral efficiency changes with delay tolerance of task, and then explain derivations.

Intuitively, if tolerable delay is very large, ($\frac{1}{\lambda_\mathrm{W}} \to \infty$), all tasks will be processed locally, and thus utilization level at SBS, $\Pi_\mathrm{A} \to 0$. On the other hand, if $\frac{1}{\lambda_\mathrm{W}} \to 0$, Π_A will attain its maximum, causing high interference and low spectral efficiency in the small cell network. So, we first set up relationship λ_W and Π_A. Then, we find the bitrate that UE at random location can achieve from nearest available SBS.

Our analysis is very similar to capacity analysis discussed in Sections 5.2.2, and 5.3. The difference is that the task is to be processed remotely, UE does not wait and connect. Instead UE uploads the task to the nearest idle cell in the network. We first find the coverage probablity, then using Shannons' capacity formula, we measure spectral efficiency in the network.

Coverage probability is already defined in Eq. (5.3). If UE connects to the nearest idle cell, probability density function of the distance to the nearest idle cell can be written as

$$f_R(r) = 2\pi\Pi_{\mathrm{I}}\rho_f r e^{-\pi\Pi_{\mathrm{I}}\rho_f r^2}. \tag{6.10}$$

In order to find Π_{I}, we rewrite the equilibrium condition in (3.9) as

$$\rho_{\mathrm{u}}\lambda_{\mathrm{u}}(1 - P_{\mathrm{R}}) = \rho_{\mathrm{c}}\Pi_{\mathrm{A}}\mu_{\mathrm{f}}, \tag{6.11}$$

where P_{R} is the proportion of locally processed task found in (6.5). Since, we do not employ energy saving strategy and focus on only spectral efficiency analysis, $\Pi_{\mathrm{S}} = 0$, and $\Pi_{\mathrm{I}} + \Pi_{\mathrm{A}} = 1$. So, Π_{I} is found. We plug Π_{I} in (6.10) and pdf of distance is found. Since UE immediately connects to the nearest idle cell, (6.10) is simplified form of (5.9). Immediate access (IA) occurs with probability one. Then, substituting (6.10) into (5.14), we have

$$\mathbb{P}(\mathrm{SINR} > \gamma) = \int_0^\infty \mathbb{P}(\mathrm{SINR} > \gamma|r) f_R(r)\mathrm{d}r, \tag{6.12}$$

where $\mathbb{P}(\mathrm{SINR} > \gamma|r)$ is equivalent as (5.15). Laplace transform (i.e. $s = \zeta\gamma r^\alpha$) in (5.15) for $\alpha = 4$ is

$$\mathcal{L}_{\mathrm{I}_{\Phi_{\mathrm{A}}}}(\zeta\gamma r^4) = e^{-p_{\mathrm{A}}\rho_{\mathrm{f}} r^2 \sqrt{\gamma}\pi^2/2}. \tag{6.13}$$

Plugging (6.13) into (5.15) gives $\mathbb{P}(\mathrm{SINR} > \gamma|r)$. Substituting $\mathbb{P}(\mathrm{SINR} > \gamma|r)$ into (6.12) completes the derivation of coverage probability.

To measure spectral efficiency, we derive the bitrate. To be more precise, we find expected value of $\log_2(1 + \mathrm{SINR})$ under the assumption that Shannon's capacity is achievable. The average bitrate if UE connects to the cell immediately can be written as:

$$\mathbb{E}\left[\log_2(1+\mathrm{SINR})\right] = \int_0^\infty f_{\mathrm{R}}(r) \int_0^\infty \frac{\ln(1+\gamma)}{\ln(2)} f_{\gamma|r}(\gamma)\mathrm{d}\gamma\mathrm{d}r, \tag{6.14}$$

where $f_{\gamma|r}(\gamma)$ is the conditional density function of SINR given that distance betwen user and transmitting cell is r. In the inner integration in (6.14), we can write conditional pdf using the tail distribution of SINR in (5.15). Then, we may write:

$$\begin{aligned} C &= \frac{1}{\ln(2)} \int_0^\infty f_{\mathrm{R}}(r) \int_0^\infty \ln(1+\gamma)\mathrm{d}\left(\mathbb{P}\left(\mathrm{SINR} > \gamma|r\right)\right)\mathrm{d}r \\ &= \frac{1}{\ln(2)} \int_0^\infty f_{\mathrm{R}}(r) \int_0^\infty \frac{\mathbb{P}\left(\mathrm{SINR} > \gamma|r\right)}{1+\gamma}\mathrm{d}\gamma\mathrm{d}r \qquad (6.15) \\ &= \frac{1}{\ln(2)} \int_0^\infty \left(\int_0^\infty \frac{\mathbb{P}\left(\mathrm{SINR} > \gamma|r\right)}{1+\gamma} f_{\mathrm{R}}(r)\mathrm{d}r\right)\mathrm{d}\gamma. \qquad (6.16) \end{aligned}$$

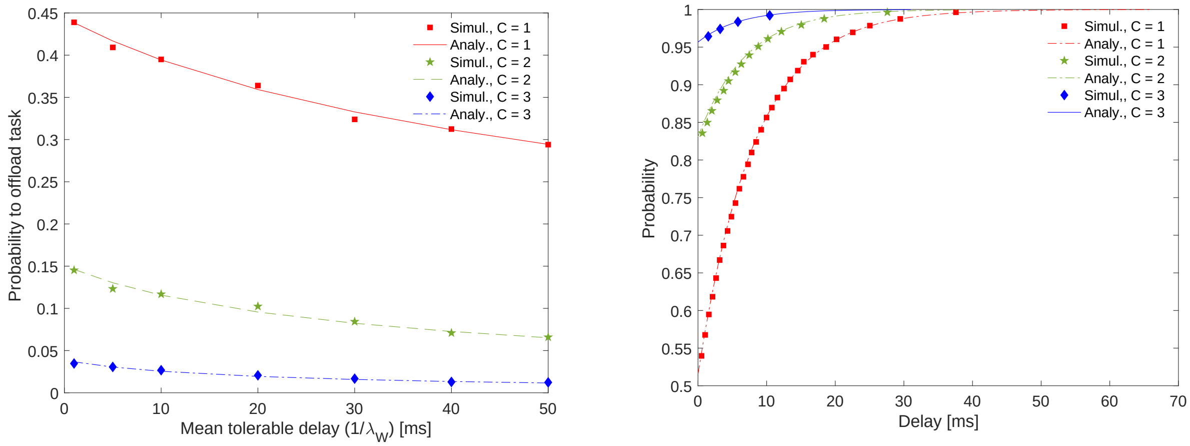
The second equality of (6.15) is due to change of variables in the inner integration. After changing the order of integration in third equality, the inner integration reduces to (6.12), and final expression becomes

$$\tau = \frac{1}{\ln(2)} \int_0^\infty \frac{\mathbb{P}\left(\mathrm{SINR} > \gamma\right)}{1+\gamma}\mathrm{d}\gamma\ . \tag{6.17}$$

6.3 Discussion

6.3.1 Verification of Distributions

In Figure 6.3, we explain probabilities analyzed in Sections 6.2.1.1 and (6.2.1.2) for $C \in \{1,2,3\}$. We assumed that UE can handle a few CPU intensive applications at a time because of its limited CPU speed. In fact, even one application such as VR can use up CPU cycles of UE. Extensive simulations are provided and analytical results are computed using (6.5) and (6.9). We observe that the analytical results nicely match the simulations for all three processing capability levels. Unlike edge processor, number of applications UE can handle, C, is small for local proces-



(a) Probability to offload

(b) Delay distribution in local CPU

Figure 6.3: (a) Probability that taks is not processed in local CPU within tolerable delay time and offloaded to edge processor. (b) Delay distribution of tasks that are computed locally for UE $1/\lambda_{\mathrm{W}} = 10$ $1/\lambda_u = 50$ ms, $1/\mu_{\mathrm{l}} = 40$ ms, number of simultaneously processed tasks $C \in \{1, 2, 3\}$.

sor due to limited CPU power, which may be entirely exhausted by high processring requirement of tasks.

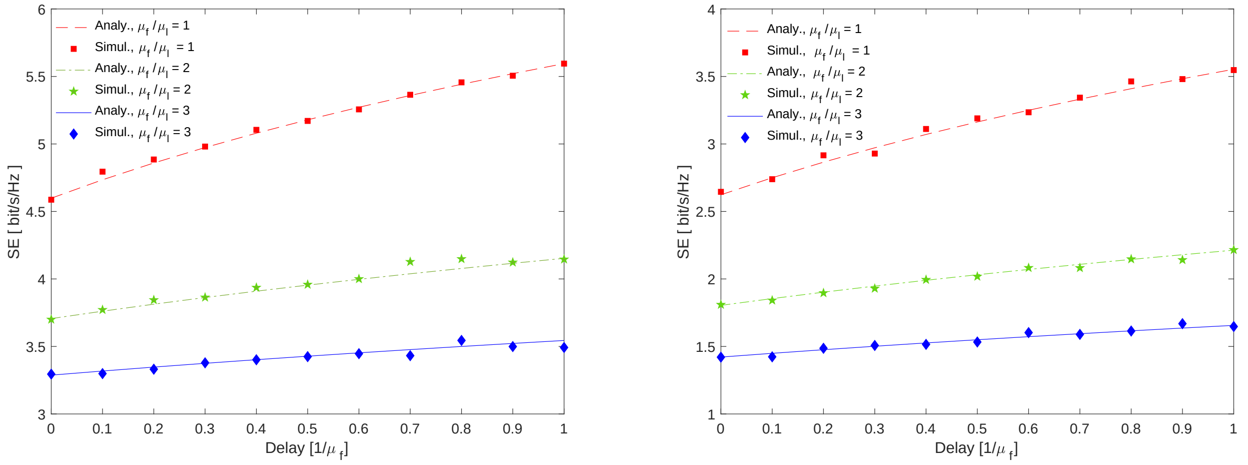
6.3.2 Delay Spectral Efficiency Tradeoff

In this part, we discuss the results of spectral efficiency analysis. We employ the same wireless channel model introduced in Section 5.1. The remaining system parameters are given in Table 6.1. We consider two cases, namely; moderate utilization and high utilization cases. In both cases, average task arrival period from UE is 100 ms. Mean end-to-end response time from edge server is 30 ms. Average local task processing period varies from 30 ms to 90 ms. These parameters corresponds network utilization of 30% and 90% (i.e. $\frac{\rho_{\mathrm{u}}\lambda_{\mathrm{u}}}{\rho_{\mathrm{c}}C\mu_{\mathrm{f}}} = \{0.3, 0.90\}$) in moderate and high utilization cases respectively. If all tasks are processed locally, utilization at the queue modeling local CPU is maximum 90% (i.e. $\frac{\rho_{\mathrm{u}}\lambda_{\mathrm{u}}}{\rho_{\mathrm{c}}C\mu_{\mathrm{l}}} = 0.90$). By choosing these parameters set, we aim to observe change in spectral efficiency with the delay tolerance.

In Figure 6.4, we observe the spectral efficiency versus delay trade-offs. Tolerable delay is normalized with end-to-end remote processing delay. We observe that spectral efficiency in low

Table 6.1: Simulation Parameters

Parameter	Value
Small cell density (ρ_c)	$0.0005\ \mathrm{m}^{-2}$
User density (ρ_u)	$\{0.0005, 0.0015\}\ \mathrm{m}^{-2}$
Task arrival rate (λ_U)	$\{0.01\}\ \mathrm{ms}^{-1}$
Mean end-to-end delay ($1/\mu_\mathrm{f}$)	30 ms
Mean local processing time ($1/\mu_\mathrm{l}$)	$(1/\mu_\mathrm{f}) \times \{1, 2, 3\}$
Mean tolerable delay ($\frac{1}{\lambda_\mathrm{W}}$)	$\left[0, \frac{1}{\mu_\mathrm{f}}\right]$ ms
Path loss exponent (α)	4



(a) Moderate network utilization

(b) High network utilization

Figure 6.4: Spectral efficiency - delay tradeoff in moderate and high networ utilizations.

utilization is higher than that of high utilization. This is due to UE being more likely to receive service from a closeby SBS. We observe in both cases that local processing is favorable when mean local processing time is nearly equal to end-to-end resoponse time ($\frac{1}{\mu_\mathrm{l}} = \frac{1}{\mu_\mathrm{f}}$). Normally, we expect that $\frac{1}{\mu_\mathrm{l}} \gg \frac{1}{\mu_\mathrm{f}}$; however, due to attenuation in mmWave channels, and scarcity of bandwidth to sustain high speed in 4G channels, end-to-end delay of remote processing may be comparable to local processing period. In that case, we observe that local processing is a good choice to improves spectral efficiency of small cell network.

6.4 Conclusions and Future Work

In this chapter, we shed light into spectral efficiency and delay tradeoff regarding the CPU intensive applications serviced by edge/fog units. We observed that if the propagation delay is large, and end to end response time is long, local processing of tasks is favorable to improve spectral efficiency of small cell networks. On the other hand, if end-to-end delay is small, local processing with delay tolerance at UE has marginal improvement on spectral efficiency of network. New protocols that make a dynamic decision between local or remote processing are necessary. Such protocols can be designed by taking into consideration of the available bandwidth, end-to-end response time, and traffic load.

Our simulation and analysis give preliminary results on spectral efficiency and delay tradeoff. Spectral efficiency-delay and energy efficiency-delay tradeoffs can be evaluated by simulations with deterministic arrival times instead of Poisson arrivals and constant delay tolerance.

Conclusions

In this thesis we offered several solutions to operate small cell networks in an energy efficient manner, and analyzed capacity-delay tradeoffs. Our main result is that when the traffic load is low, delay tolerant traffic of user equipment can be utilized to design energy saving schemes, and improving spectral efficiency. In dense deployments, simple traffic load based algorithms can decrease small cell energy consumption significantly. However, it can be further decreased by optimizing sleep time with respect to traffic load, user and SBS densities.

In Chapter 3, energy-efficient operation of small cells is treated as a Markov chain, which is then used to analyze the associated energy savings. In particular, a simple delayed access scheme is introduced for user equipment. Subsequently, the thesis investigates various aspects of delay such as delay-average power consumption, delay-transmission power relationships.

Chapter 4 introduces a novel and practical traffic load metric for SCN. Statistical properties of the load metric are extracted by finding its fitting distribution. Then, the load metric is used to design centralized and distributed energy saving schemes. The performances of these schemes are evaluated by comparing two benchmark algorithms in low and mid traffic periods.

Chapter 5 analyses the impact of UE's delayed access on an SCN's coverage probability and bitrate. Assuming that the locations of SBSs' follow a homogeneous Poisson process, coverage probability and capacity are derived using stochastic geometry methods. The analysis further allows us to obtain optimal transmission range of UE that maximizes coverage probability with respect to given delay tolerance.

Chapter 6 investigates delay spectral efficiency trade-off in edge applications. Local processing delay is modelled as an M/M/c queue with impatience and statistical properties of the queue such as distribution and mean of queuing delay, are analyzed. Tasks in the queue that are not processed within delay budget time are uploaded to edge server. Analysis of the queue allows us to measure relationship between network utilization and tolerable delay of the task. After network utilization is obtained, further analysis is made to obtain bitrate; showing the spectral efficiency-delay tradeoff.

References

[1] *LTE Release 12 and Beyond*, White Paper, Nokia Siemens Networks, Oct. 2012.

[2] Q. Staff, "Rising to meet the 1000x mobile data challenge," *QUALCOMM*, 2012.

[3] T. Nakamura, S. Nagata, A. Benjebbour, Y. Kishiyama, T. Hai, S. Xiaodong, Y. Ning, and L. Nan, "Trends in small cell enhancements in lte advanced," *IEEE Commun. Mag.*, vol. 51, no. 2, pp. 98–105, 2013.

[4] S. Carlaw and C. Wheelock, "Femtocell market challenges and opportunities: Cellular-based fixed mobile convergence for consumers, smes, and enterprises," *ABI Research*, 2007.

[5] C. Bouras, V. Kokkinos, and A. Papazois, "Financing and pricing small cells in next-generation mobile networks," in *Int. Conf, on Wired/Wireless Internet Communs.*, Springer, 2014, pp. 41–54.

[6] G. I. Gartner and I. Green, "The new industry shockwave, presentation at symposium," in *ITXPO conference, April*, 2007.

[7] M. Webb *et al.*, "Smart 2020: Enabling the low carbon economy in the information age," *The Climate Group. London*, vol. 1, no. 1, pp. 1–1, 2008.

[8] *Greentouch*, http://www.greentouch.org/index.php?page=about-us, Accessed: 2010-09-30.

[9] *ICT EARTH*, https://www.ict-earth.eu, Accessed: 2010-09-30.

[10] *Greenet*, http://www.fp7-greenet.eu, Accessed: 2010-09-30.

[11] P. Jonsson, S. Bävertoft, R. Möller, H. Andersson, M. Björn, S. Carson, S. Frost, I. Godor, M. Halen, P. Kersch, P. Lindberg, V. Singh, and L. Wieweg, "Ericsson mobility report: On the pulse of the networked society," *Ericsson*, 2014.

[12] I. Andone, K. Blaszkiewicz, M. Eibes, B. Trendafilov, C. Montag, and A. Markowetz, "How age and gender affect smartphone usage," in *Proc. of the 2016 ACM Int. Joint Conf. on Pervasive and Ubiquitous Computing: Adjunct*, ser. UbiComp '16, Heidelberg, Germany: ACM, 2016, pp. 9–12, ISBN: 978-1-4503-4462-3.

[13] S. Zhao, J. Ramos, J. Tao, Z. Jiang, S. Li, Z. Wu, G. Pan, and A. K. Dey, "Discovering different kinds of smartphone users through their application usage behaviors," in *Proc.*

of the 2016 ACM Int. Joint Conf. on Pervasive and Ubiquitous Computing, ser. UbiComp '16, Heidelberg, Germany: ACM, 2016, pp. 498–509, ISBN: 978-1-4503-4461-6. DOI: 10.1145/2971648.2971696.

[14] M. Ajmone Marsan, L. Chiaraviglio, D. Ciullo, and M. Meo, "Optimal energy savings in cellular access networks," in *Int. Conf. on Commun. Workshops*, 2009, pp. 1–5. DOI: 10.1109/ICCW.2009.5208045.

[15] 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, 2013, ISSN: 1536-1276. DOI: 10.1109/TWC.2013.032013.120494.

[16] T. Eyers and H. Schulzrinne, "Predicting internet telephony call setup delay," in *Proc. 1st IP-Telephony Workshop*, 2000.

[17] *Common voip service quality thresholds*, Tektronix. [Online]. Available: http://www.tek.com/dl/VoIPServiceQualityMetricsandThresholdsPoster.pdf.

[18] D. F. Galletta, R. Henry, S. McCoy, and P. Polak, "Web site delays: How tolerant are users?" *Journal of the Association for Information Systems*, vol. 5, no. 1, p. 1, 2004.

[19] D. Kirkpatrick. (2016). Google: 53% of mobile users abandon sites that take over 3 seconds to load, [Online]. Available: https://www.marketingdive.com/news/google-53-of-mobile-users-abandon-sites-that-take-over-3-seconds-to-load/426070/ (visited on 09/10/2019).

[20] L. Meng, S. Liu, and A. D. Striegel, "Characterizing the utility of smartphone background traffic," in *Proc. 23rd Int Conf. on Comput. Commun. and Networks (ICCCN)*, 2014, pp. 1–5.

[21] J. Huang, F. Qian, Z. M. Mao, S. Sen, and O. Spatscheck, "Screen-off traffic characterization and optimization in 3g/4g networks," in *Proc. of ACM Internet Measurement Conf.*, ser. IMC '12, Boston, Massachusetts, USA: ACM, 2012, pp. 357–364, ISBN: 978-1-4503-1705-4. DOI: 10.1145/2398776.2398813.

[22] Y. Moon, D. Kim, Y. Go, Y. Kim, Y. Yi, S. Chong, and K. Park, "Cedos: A network architecture and programming abstraction for delay-tolerant mobile apps," *IEEE/ACM Trans. Netw.*, vol. 25, no. 2, pp. 646–661, 2017.

[23] H. S. Dhillon, "Fundamentals of heterogeneous cellular networks," PhD thesis, 2013.

[24] F. Zafari, A. Gkelias, and K. K. Leung, "A survey of indoor localization systems and technologies," *CoRR*, vol. abs/1709.01015, 2017. arXiv: 1709.01015. [Online]. Available: http://arxiv.org/abs/1709.01015.

[25] H. Wang, S. Sen, A. Elgohary, M. Farid, M. Youssef, and R. R. Choudhury, “No need to war-drive: Unsupervised indoor localization,” in *Proc. of ACM MobiSys*, ser. MobiSys ’12, Low Wood Bay, Lake District, UK: ACM, 2012, pp. 197–210, ISBN: 978-1-4503-1301-8.

[26] Y. Kinoshita, T. Tsuchiya, and S. Ohnuki, “Frequency common use between indoor and urban cellular radio-research on frequency channel doubly reused cellular system,” in *Proc. IEEE 39th Veh. Technol. Conf.*, 1989, 329–335 vol.1.

[27] M. I. Silventoinen, M. Kuusela, and P. A. Ranta, “Analysis of a new channel access method for home base station,” in *Proc. Int. Conf. Universal Personal Commun. (ICUPC)*, vol. 2, 1996, 930–935 vol.2.

[28] 3GPP, “Technical Specification Group Radio Access Network; UTRAN Iuh Interface HNBAP signalling (Release 8),” 3rd Generation Partnership Project (3GPP), Technical Specification (TS) 25.469, Nov. 2008, Version V1.0.0. [Online]. Available: https://portal.3gpp.org/desktopmodules/Specifications/SpecificationDetails.aspx?specificationId=1222.

[29] 3GPP, “3rd Generation Partnership Project; Technical Specification Group Services and System Aspects; Service requirements for Home Node B (HNB) and Home eNode B (HeNB) (Release 15),” 3rd Generation Partnership Project (3GPP), Technical Specification (TS) 22.220, Jul. 2019, Version V15.0.0. [Online]. Available: https://www.3gpp.org/ftp/Specs/archive/22_series/22.220/22220-f00.zip.

[30] 3GPP, “3rd Generation Partnership Project; Technical Specification Group Radio Access Network; Scenarios and requirements for small cell enhancements for E-UTRA and E-UTRAN (Release 15),” 3rd Generation Partnership Project (3GPP), Technical Specification (TS) 36.932, Jun. 2018, Version V15.0.0. [Online]. Available: https://www.3gpp.org/ftp/Specs/archive/36_series/36.932/36932-f00.zip.

[31] FUJITSU Network Communications Inc. (2019). High-Capacity Indoor Wireless Solutions: Picocell or Femtocell? Accessed: 2019-09-30, [Online]. Available: https://www.fujitsu.com/us/Images/High-Capacity-Indoor-Wireless.pdf.

[32] Small Cell Forum, “Small cells, what’s the big idea?” 2014.

[33] Small Cell Forum, “5G FAPI: PHY API Specification (Relase 10),” Tech. Rep. 222.10.01, Jun. 2019. [Online]. Available: https://www.smallcellforum.org/5g-phy-api-release/.

[34] Huawei. (Feb. 2016). Five Trends to Small Cell 2020, [Online]. Available: http://www-file.huawei.com/~/media/CORPORATE/PDF/News/Five-Trends-To-Small-Cell-2020-en.pdf.

[35] T. Chen, H. Kim, and Y. Yang, “Energy efficiency metrics for green wireless communications,” in *Proc. Int. Conf. Wireless Commun. and Signal Process. (WCSP)*, IEEE, 2010, pp. 1–6.

[36] J. Zhang, Y. Ji, X. Xu, H. Li, Y. Zhao, and J. Zhang, “Energy efficient baseband unit aggregation in cloud radio and optical access networks,” *IEEE J. Opt. Commun. Netw.*, vol. 8, no. 11, pp. 893–901, 2016.

[37] 3GPP, “Home eNode B radio frequency requirements analysis,” 3rd Generation Partnership Project (3GPP), TR 36.921, Apr. 2010. [Online]. Available: http://www.3gpp.org/DynaReport/36921.htm.

[38] A. Vasjanov and V. Barzdenas, “A Review of Advanced CMOS RF Power Amplifier Architecture Trends for Low Power 5G Wireless Networks,” *Electronics*, vol. 7, no. 11, 2018, ISSN: 2079-9292. [Online]. Available: https://www.mdpi.com/2079-9292/7/11/271.

[39] H. Claussen, I. Ashraf, and L. T. Ho, “Dynamic idle mode procedures for femtocells,” *Bell Labs Technical Journal*, vol. 15, no. 2, pp. 95–116, 2010.

[40] 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?” *Wireless Commun.*, vol. 18, no. 5, pp. 40–49, 2011.

[41] M. Deruyck, D. De Vulder, W. Joseph, and L. Martens, “Modelling the power consumption in femtocell networks,” in *Proc. IEEE Wireless Commun and Networking Conf. Workshops (WCNCW)*, 2012, pp. 30–35.

[42] 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, 2016, ISSN: 0018-9545. DOI: 10.1109/TVT.2015.2413382.

[43] Y. Hou and D. I. Laurenson, “Energy efficiency of high qos heterogeneous wireless communication network,” in *Proc. IEEE VTC Fall’10*, IEEE, 2010, pp. 1–5.

[44] B. Partov, D. J. Leith, and R. Razavi, “Energy-aware configuration of small cell networks,” in *Proc. IEEE 25th Annual Int. Symp. on Personal, Indoor, and Mobile Radio Commun. (PIMRC)*, 2014, pp. 1403–1408.

[45] Q. Zhu, X. Wang, and Z. Qian, “Energy-efficient small cell cooperation in ultra-dense heterogeneous networks,” *IEEE Commun. Lett.*, vol. 23, no. 9, pp. 1648–1651, 2019.

[46] M. Oikonomakou, A. Antonopoulos, L. Alonso, and C. Verikoukis, “Energy sharing and trading in multi-operator heterogeneous network deployments,” *IEEE Trans. Veh. Technol.*, vol. 68, no. 5, pp. 4975–4988, 2019. DOI: 10.1109/TVT.2019.2902925.

[47] W. Vereecken, L 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. WONS*, IEEE, 2012, pp. 63–66.

[48] M. Ismail and W. Zhuang, “Network cooperation for energy saving in green radio communications,” *Wireless Commun.*, vol. 18, no. 5, pp. 76–81, 2011.

[49] T. Yang and L. Zhang, “Approaches to enhancing autonomous power control at femto under co-channel deployment of macrocell and femtocell,” in *Proc. IEEE PIMRC*, IEEE, 2011, pp. 71–75.

[50] D. Shin and S. Choi, “Dynamic power control for balanced data traffic with coverage in femtocell networks,” in *Proc. IEEE IWCMC*, IEEE, 2012, pp. 648–653.

[51] M. Pavel and B. Zdenek, “QoS-guaranteed power control mechanism based on the frame utilization for femtocells,” *EURASIP Journal on Wireless Communications and Networking*, vol. 2011, 2011.

[52] C. S. Chen, F. Baccelli, and L. Roullet, “Joint optimization of radio resources in small and macro cell networks,” in *Proc. IEEE Veh. Technol. Conf.*, IEEE, 2011, pp. 1–5.

[53] K. I. Pedersen, Y. Wang, S. Strzyz, and F. Frederiksen, “Enhanced inter-cell interference coordination in co-channel multi-layer LTE-advanced networks,” *IEEE Trans. Wireless Commun.*, vol. 20, no. 3, pp. 120–127, 2013.

[54] C. Li, J. Zhang, and K. B. Letaief, “User-centric intercell interference coordination in small cell networks,” in *Proc. IEEE Int. Conf. Commun. (ICC)*, 2014, pp. 5747–5752.

[55] B. Soret, K. I. Pedersen, N. T. K. Jørgensen, and V. Fernández-López, “Interference coordination for dense wireless networks,” *IEEE Commun. Mag.*, vol. 53, no. 1, pp. 102–109, 2015.

[56] C.-C. Kuan, G.-Y. Lin, and H.-Y. Wei, “Energy Efficient Networking with IEEE 802.16 m Femtocell Low Duty Mode,” *Mobile Networks and Applications*, vol. 17, no. 5, pp. 674–684, 2012.

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

[58] W. Vereecken, M. Deruyck, D. Colle, W. Joseph, M. Pickavet, L. Martens, and P. Demeester, “Evaluation of the potential for energy saving in macrocell and femtocell networks using a heuristic introducing sleep modes in base stations,” *EURASIP Journal on Wireless Communications and Networking*, vol. 2012, no. 1, pp. 1–14, 2012.

[59] Z. Huang, H. Xia, Z. Zeng, and Y. Liu, “Optimization of energy efficiency for ofdma femtocell networks based on effective capacity,” in *Proc. IEEE Veh. Technol. Conf.*, IEEE, 2012, pp. 1–5.

[60] Dapeng Wu and R. Negi, “Effective capacity: a wireless link model for support of quality of service,” *IEEE Trans. Commun.*, vol. 2, no. 4, pp. 630–643, 2003.

[61] 3GPP, “LTE;Evolved Universal Terrestrial Radio Access (E-UTRA);Potential solutions for energy saving for E-UTRAN,” 3rd Generation Partnership Project (3GPP), Technical Specification (TS) 36.927, Jul. 2011, Version V10.0.0. [Online]. Available: https://www.etsi.org/deliver/etsi_tr/136900_136999/136927/10.00.00_60/tr_136927v100000p.pdf.

[62] Y. Li, H. Celebi, M. Daneshmand, C. Wang, and W. Zhao, “Energy-efficient femtocell networks: Challenges and opportunities,” *IEEE Wireless Commun. Mag.*, vol. 20, no. 6, pp. 99–105, 2013.

[63] T. Q. S. Quek, W. C. Cheung, and M. Kountouris, “Energy efficiency analysis of two-tier heterogeneous networks,” in *17th European Wireless 2011 - Sustainable Wireless Technologies*, 2011, pp. 1–5.

[64] H. Klessig, A. J. Fehske, and G. P. Fettweis, “Energy efficiency gains in interference-limited heterogeneous cellular mobile radio networks with random micro site deployment,” in *Proc. 34th IEEE Sarnoff Symp.*, 2011, pp. 1–6.

[65] S. Wang, W. Guo, and T. O’Farrell, “Low energy indoor network: Deployment optimisation,” *EURASIP Journal on Wireless Communications and Networking*, vol. 2012, no. 1, p. 193, 2012, ISSN: 1687-1499. DOI: 10.1186/1687-1499-2012-193. [Online]. Available: https://doi.org/10.1186/1687-1499-2012-193.

[66] J. G. Andrews, H. Claussen, M. Dohler, S. Rangan, and M. C. Reed, “Femtocells: Past, Present, and Future,” *IEEE J. Sel. Areas Commun.*, vol. 30, no. 3, pp. 497–508, 2012.

[67] T. Rappaport, *Wireless Communications: Principles and Practice*, 2nd. Upper Saddle River, NJ, USA: Prentice Hall PTR, 2001, ISBN: 0130422320.

[68] A. Al-Hourani, R. J. Evans, and S. Kandeepan, “Nearest Neighbor Distance Distribution in Hard-Core Point Processes,” *IEEE Commun. Lett.*, vol. 20, no. 9, pp. 1872–1875, 2016.

[69] Z. Khalid and S. Durrani, “Distance Distributions in Regular Polygons,” *IEEE Trans. Veh. Technol.*, vol. 62, no. 5, pp. 2363–2368, 2013.

[70] C. Liu and L. Wang, “Optimal cell load and throughput in green small cell networks with generalized cell association,” *IEEE J. Sel. Areas Commun.*, vol. 34, no. 5, pp. 1058–1072, 2016.

[71] A. J. Mahbas, H. Zhu, and J. Wang, “Impact of Small Cells Overlapping on Mobility Management,” *IEEE Trans. Wireless Commun.*, vol. 18, no. 2, pp. 1054–1068, 2019.

[72] J. Yoon and G. Hwang, “Distance-based inter-cell interference coordination in small cell networks: Stochastic geometry modeling and analysis,” *IEEE Trans. on Wireless Commun.*, vol. 17, no. 6, pp. 4089–4103, 2018.

[73] M. L. Moss and C. Qing, “The dynamic population of manhattan,” 2012. [Online]. Available: http://wagner.nyu.edu/rudincenter/publications/dynamic_pop_manhattan.pdf.

[74] 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, 2011.

[75] Z. Lin, Y. Gao, B. Gong, X. Zhang, and D. Yang, “Stochastic geometry study on small cell on/off adaptation,” in *Proc. Int. Conf. Commun. Netw. in China*, 2014, pp. 331–334.

[76] H. Celebi, N. Maxemchuk, Y. Li, and I. Guvenc, “Energy reduction in small cell networks by a random on/off strategy,” in *Proc. of GLOBECOM Workshop*, 2013, pp. 176–181.

[77] L. Accessory Fulfillment Center. (). Small Cell Installations: Microcell, Metrocell, Picocell, Femtocell for AT&T, Verizon, T-Mobile, Sprint Coverage, [Online]. Available: https://www.signalbooster.com/pages/small-cell-installations-microcell-metrocell-picocell-femtocell (visited on 2019).

[78] RepeaterStore. (). Femtocells, Microcells, and Metrocells: The Complete Guide to Small Cells, [Online]. Available: https://www.repeaterstore.com/pages/femtocell-and-microcell (visited on 2019).

[79] A. Fanghänel, S. Geulen, M. Hoefer, and B. Vöcking, “Online Capacity Maximization in Wireless Networks,” in *Proc. of the Twenty-second Annu. ACM Symp. on Parallelism in Algorithms and Architectures*, ser. SPAA ’10, Greece: ACM, 2010, pp. 92–99, ISBN: 978-1-4503-0079-7.

[80] P. Bremaud, J. E. Marsden, L. Sirovic, and W. Jager, Eds., *Markov chains: Gibbs Fields, Monte Carlo Simulation, and Queues*. New York: Springer-Verlag, 1999.

[81] H. Claussen, I. Ashraf, and L. T. Ho, “Dynamic idle mode procedures for femtocells,” *Bell Labs Technical Journal*, vol. 15, no. 2, pp. 95–116, 2010.

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

[83] T. Q. Quek, G. de la Roche,
bibinitperiodI. Güvenç, and M. Kountouris, *Small cell networks: Deployment, PHY techniques, and resource management*. Cambridge University Press, 2013.

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

[85] S. Samarakoon, M. Bennis, W. Saad, M. Debbah, and M. Latva-aho, “Ultra dense small cell networks: Turning density into energy efficiency,” *IEEE J. Select. Areas Commun. (JSAC)*, vol. 34, no. 5, pp. 1267–1280, 2016.

[86] I. Siomina and D. Yuan, “Analysis of Cell Load Coupling for LTE Network Planning and Optimization,” *IEEE Trans. Wireless Commun.*, vol. 11, no. 6, pp. 2287–2297, 2012, ISSN: 1536-1276.

[87] A. J. Fehske and G. P. Fettweis, “Aggregation of variables in load models for interference-coupled cellular data networks,” in *Proc. Int. Conf. Commun. (ICC)*, 2012, pp. 5102–5107. DOI: 10.1109/ICC.2012.6363999.

[88] H. Klessig, A. Fehske, G. Fettweis, and J. Voigt, “Cell load-aware energy saving management in self-organizing networks,” in *2013 IEEE 78th Vehicular Technology Conference (VTC Fall)*, 2013, pp. 1–6.

[89] C. Li, J. Zhang, and K. B. Letaief, “Throughput and energy efficiency analysis of small cell networks with multi-antenna base stations,” *IEEE Trans. Wireless Commun.*, vol. 13, no. 5, pp. 2505–2517, 2014.

[90] Y. S. Soh, T. Q. Quek, M. Kountouris, and H. Shin, “Energy efficient heterogeneous cellular networks,” *IEEE J. Select. Areas Commun.*, vol. 31, no. 5, pp. 840–850, 2013.

[91] A Merwaday and I Güvenç, “Optimisation of FeICIC for energy efficiency and spectrum efficiency in LTE-advanced HetNets,” *IET Electronics Letters*, vol. 52, no. 11, pp. 982–984, 2016.

[92] C. Peng, S.-B. Lee, S. Lu, and H. Luo, “GreenBSN: Enabling energy-proportional cellular base station networks,” *IEEE Trans. Mobile Comput.*, vol. 13, no. 11, pp. 2537–2551, 2014, ISSN: 1536-1233.

[93] Y. Zeng, K. Xiang, D. Li, and A. V. Vasilakos, “Directional routing and scheduling for green vehicular delay tolerant networks,” *Wireless Networks*, vol. 19, no. 2, pp. 161–173, 2013.

[94] S. He, X. Li, J. Chen, P. Cheng, Y. Sun, and D. Simplot-Ryl, “EMD: Energy-efficient P2P message dissemination in delay-tolerant wireless sensor and actor networks,” *IEEE J. Select. Areas Commun. (JSAC)*, vol. 31, no. 9, pp. 75–84, 2013.

[95] Y. Cao and Z. Sun, “Routing in delay/disruption tolerant networks: A taxonomy, survey and challenges,” *IEEE Commun. Surveys & Tutorials*, vol. 15, no. 2, pp. 654–677, 2013.

[96] X. Guo, Z. Niu, S. Zhou, and P. R. Kumar, “Delay-constrained energy-optimal base station sleeping control,” *IEEE J. Sel. Areas Commun.*, vol. 34, no. 5, pp. 1073–1085, 2016.

[97] H. Celebi and I. Guvenc, “Load analysis and sleep mode optimization for energy-efficient 5G small cell networks,” in *Proc. IEEE Int. Conf. Commun. (ICC) Workshops*, 2017, pp. 1159–1164.

[98] H. Çelebi, Y. Yapıcı, I. Guvenc, and H. Schulzrinne, “Load-Based On/Off Scheduling for Energy-Efficient Delay-Tolerant 5G Networks,” *IEEE Trans. on Green Commun. and Networking*, vol. 3, no. 4, pp. 955–970, 2019, ISSN: 2473-2400.

[99] I. Ashraf, F. Boccardi, and L. Ho, “Sleep mode techniques for small cell deployments,” *IEEE Commun. Mag.*, vol. 49, no. 8, pp. 72–79, 2011, ISSN: 0163-6804.

[100] 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. Annu. Conf. Wireless On-Demand Netw. Syst. and Services (WONS)*, 2012, pp. 63–66.

[101] M. Tanemura, “Statistical distributions of poisson voronoi cells in two and three dimensions,” *FORMA*, vol. 18, no. 4, pp. 221–247, 2003.

[102] J.-S. Ferenc and Z. Néda, “On the size distribution of poisson voronoi cells,” *Physica A: Statistical Mechanics and its Applications*, vol. 385, no. 2, pp. 518–526, 2007.

[103] C. Davies, “Size distribution of atmospheric particles,” *Journal of Aerosol Science*, vol. 5, no. 3, pp. 293–300, 1974.

[104] S. M. Ross, *Introduction to Probability Models*, 10th. Academic Press, 2009.

[105] D. Milios, “Probability distributions as program variables,” *Master’s thesis, School of Informatics, University of Edinburgh*, 2009.

[106] M. Fewell, “Area of common overlap of three circles,” Defence Science and Technology Organisation, Report DSTO-TN-0722, Oct. 2006.

[107] A. Blogowski, O. Klopfenstein, and B. Renard, “Dimensioning X2 backhaul link in LTE networks,” in *Proc. IEEE Int. Conf. Commun. (ICC)*, 2012, pp. 2768–2773.

[108] J. Xu, J. Zhang, and J. G. Andrews, “On the accuracy of the wyner model in cellular networks,” *IEEE Trans. Wireless Commun.*, vol. 10, no. 9, pp. 3098–3109, 2011.

[109] 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, 2016, ISSN: 0018-9545. DOI: 10.1109/TVT.2015.2413382.

[110] C. Liu and M. Bennis, “Ultra-reliable and low-latency vehicular transmission: An extreme value theory approach,” *IEEE Commun. Lett.*, vol. 22, no. 6, pp. 1292–1295, 2018.

[111] S. Samarakoon, M. Bennis, W. Saad, and M. Debbah, “Federated learning for ultra-reliable low-latency V2V communications,” *CoRR*, vol. abs/1805.09253, 2018. arXiv: 1805.09253. [Online]. Available: http://arxiv.org/abs/1805.09253.

[112] H. Celebi, I. Guvenc, and H. Schulzrinne, “Capacity and energy-efficiency of delayed access scheme for small cell networks,” in *Proc. of WTS*, 2019, pp. 1–6.

[113] S. N. Chiu, D. Stoyan, W. S. Kendall, and J. Mecke, *Stochastic geometry and its applications*. John Wiley & Sons, 2013.

[114] J. Pearce. (2017). VR and AR device shipments to hit 99m by 2021, [Online]. Available: https://www.capacitymedia.com/articles/3755961/VR-and-AR-device-shipments-to-hit-99m-by-2021 (visited on 10/02/2017).

[115] T. Merel. (2017). The reality of VR/AR growth, [Online]. Available: https://techcrunch.com/2017/01/11/the-reality-of-vrar-growth/ (visited on 2017).

[116] L. Liu, R. Zhong, W. Zhang, Y. Liu, J. Zhang, L. Zhang, and M. Gruteser, “Cutting the cord: Designing a high-quality untethered vr system with low latency remote rendering,” in *Proc. ACM MobiSys ’18*, ser. MobiSys ’18, Munich, Germany: ACM, 2018, pp. 68–80, ISBN: 978-1-4503-5720-3.

[117] Cisco and its affiliates. (2015). Fog computing and the internet of things: Extend the cloud to where the things are, [Online]. Available: https://www.cisco.com/c/dam/en_us/solutions/trends/iot/docs/computing-overview.pdf (visited on 2015).

[118] L. A. Shay. (Jun. 2019). Embraer Launches Big Data Analytics Platform, [Online]. Available: https://www.mro-network.com/emerging-technology/embraer-launches-big-data-analytics-platform.

[119] Y. Wen, W. Zhang, and H. Luo, “Energy-optimal mobile application execution: Taming resource-poor mobile devices with cloud clones,” in *Proc. IEEE INFOCOM*, 2012, pp. 2716–2720.

[120] M. S. Elbamby, C. Perfecto, M. Bennis, and K. Doppler, “Toward low-latency and ultra-reliable virtual reality,” *IEEE Netw.*, vol. 32, no. 2, pp. 78–84, 2018, ISSN: 0890-8044.

[121] M. S. Elbamby, M. Bennis, W. Saad, M. Latva-aho, and C. S. Hong, “Proactive edge computing in fog networks with latency and reliability guarantees,” *EURASIP Journal on Wireless Communications and Networking*, vol. 2018, no. 1, p. 209, 2018, ISSN: 1687-1499.

[122] F. Qian, L. Ji, B. Han, and V. Gopalakrishnan, “Optimizing 360 video delivery over cellular networks,” in *Proc. of the 5th Workshop on All Things Cellular: Operations, Applications and Challenges*, ser. ATC ’16, New York City, New York: ACM, 2016, pp. 1–6, ISBN: 978-1-4503-4249-0. DOI: 10.1145/2980055.2980056.

[123] M. Bibinger, “Notes on the sum and maximum of independent exponentially distributed random variables with different scale parameters,” *arXiv preprint arXiv:1307.3945*, 2013.