2018

Optimal resource scheduling for energy-efficient next generation wireless networks

Taewoon Kim
Iowa State University

Optimal resource scheduling for energy-efficient next generation wireless networks

by

Taewoon Kim

A dissertation submitted to the graduate faculty
in partial fulfillment of the requirements for the degree of
DOCTOR OF PHILOSOPHY

Major: Computer Engineering

Program of Study Committee:
Daji Qiao, Co-major Professor
Jien (Morris) Chang, Co-major Professor
Phillip H. Jones
Ying Cai
Yong Guan

The student author, whose presentation of the scholarship herein was approved by the program of study committee, is solely responsible for the content of this dissertation. The Graduate College will ensure this dissertation is globally accessible and will not permit alterations after a degree is conferred.

Iowa State University

Ames, Iowa

2018

DEDICATION

To my family

TABLE OF CONTENTS

Page

LIST OF TABLES

LIST OF FIGURES

ACKNOWLEDGEMENTS

I would like to take this opportunity to express my thanks to those who guided and helped me with various aspects of conducting research and the writing of this dissertation. First and foremost, I would like to thank Dr. Daji Qiao and Dr. Jien Morris Chang for their guidance, patience and support throughout my graduate study. Their insights and words of encouragement have often inspired me and renewed my hopes for completing my graduate education. I would like to thank my committee members, Dr. Phillip H. Jones, Dr. Ying Cai and Dr. Yong Guan, for their efforts and contributions to this work. I would also like to thank Dr. Ahmed E. Kamal for his guidance during the initial stage of my research. I would like to thank my labmates, Wenhao, Danny, Chris, Chan-Ching, Yu-Wen, Mohammad, Di, John and Zhiming and my friends, Min Sang, Sammy, Huiwon, Kyungwon and Dongwook. They made the atmosphere much more fun and I will miss all the joyful memories we created together. Finally, I would like to express my sincere appreciation to my family for their endless and unconditional support throughout my life.

ABSTRACT

Cellular networks can provide highly available and reliable communication links to the Internet of Things (IoT) applications, letting the connected Things paradigm gain much more momentum than ever. Also, the rich information collected from the Things with sensing capabilities can guide the network operator to an unforeseen direction, allowing the underlying cellular networks to be further optimized. In this regard, the cellular networks and IoT are conceived as the key components of the beyond-4G and future 5G networks. Therefore, in this dissertation, we study each of the two components in depth, focusing on how to optimize the networking resources for the quality service and better energy-efficiency. To begin with, we study the heterogeneous cellular network architecture which is a major enhancement to the current 4G network by means of the base station (BS) densification and traffic offloading. In particular, the densely deployed short-range, low-power smallcell base stations (SBSs) can significantly improve the frequency reuse, throughput performance and the energy-efficiency. We then study the heterogeneous C-RAN (cloud radio access network), which is one of the core enablers of the next generation 5G cellular networks. In particular, with the high availability provided by the long-range macro BS (MBS), the heterogeneous C-RAN (H-CRAN) can effectively enhance the overall resource utilization compared to the conventional C-RANs. In each study, we propose an optimal resource scheduling and service provisioning scheme to provide a quality service to users in a resource-efficient manner. In addition, we carry out two studies for the Internet of Things (IoT) networks operating with the IEEE 802.11ah standard. Specifically, we introduce energy-efficient device management algorithms for the battery-operated, resource-constrained IoT sensor devices to prolong their lifetime by optimally scheduling their activation. The enhanced power saving mechanism and the optimal sensing algorithm that we propose in each study can effectively improve both the energy-efficiency of the IoT devices and the lifetime of the entire network.

CHAPTER 1. OVERVIEW

In this chapter, we introduce the background and motivation of the studies to be discussed in this dissertation. Then, the overall objective of the research is addressed, followed by the organization of the dissertation.

1.1 Introduction

Internet-accessible mobile handheld devices have changed our daily lives to a large extent, which can be easily seen by the unprecedented increase in the data usage from mobile devices. In the United States, for example, the Internet usage from mobile applications exceeded that from desktop computers in 2014 [1]. Furthermore, in 2016, it was reported that the Internet usage on mobile devices surpassed that on desktop computers worldwide [2]. Such a massive traffic shift from desktop computers to mobile devices has been driven by many factors, such as widespread use of the Internet-accessible mobile devices and a surge of bandwidth-hungry applications. In addition, the technologies allowing devices to stay connected anytime/anywhere, such as smart home/office, ambient assisted living, and Internet of Things (IoT), even expedited such a transition.

As a result, the Internet traffic over wireless networks has been sharply increased, and it is estimated that by 2020 it will exceed 500 exabytes [3]. Similarly, as pointed out in [4] the total amount of traffic to be generated by mobile devices is expected to be increased by tenfold between 2014 and 2019. It is noteworthy that these predictions in the growing data usage also imply a significant increase in the number of the Internet-accessible devices. As a result, it is inevitable that the current 4G network and its ecosystem will be saturated in a very near future despite all the efforts that have been made to extend its capacity to the theoretical limit. Therefore, it is essential to transition to the next generation networking system to accommodate not only the ever-increasing traffic demand with short delay and high reliability, but also the gigantic number of Internet-accessible devices to be deployed everywhere.

To this end, there has been a series of movements to make a step towards the fifth-generation (5G) networking system. The key requirements for 5G are 1000x aggregate data rate increase compared to 4G, a round-trip latency of approximately 1 ms, reduced energy consumption, 99.9% of reliability, massive number of connections (i.e., 1,000,000 devices/km^2 of connection density), to name a few [3] [5]. Although it still is challenging, both academia and industry have been striving with a tremendous amount of efforts to realize the 5G networking system in the near future, and they indeed have seen some notable progress on it.

As summarized and classified in [3] one of the key enablers for the 5G networking system is the heterogeneous architecture and the densification of the low-power, short-range base stations (BSs). The heterogeneous network configuration can greatly enhance the network performance. Densely deployed short-range low-power base stations (i.e., smallcell base stations or SBS) can be used to offload users' traffic from the conventional macro base station (MBS), preventing MBSs from being congested. By taking advantage of the shorter distance to the users, the low-power SBSs can achieve high throughput performance with low energy consumption, increasing the area spectral efficiency and energy-efficiency. In addition, for being low-cost and easy to deploy/operate, SBSs can reduce both the capital and operational expenditures.

Besides the densification of the low-power BSs, the power of cloud computing and its success in the market has resulted in the new design of the cellular networks, called cloud radio access network (C-RAN). C-RAN, which is conceived as one of the core enablers for the 5G networking system, is composed of three major components: RRH (remote radio head), fronthaul network, and BBU (baseband processing unit) pool. The MAC and PHY functionalities are shifted from the conventional BSs to the cloud computing platform (i.e., BBU pool) to take advantage of its powerful processing capability. Also, such a shift in functionality makes the RRHs low-complex and low-cost, enabling dense deployment of RRHs with an affordable capital expenditure. Also, the centrality of information processing at the BBU pool allows an easier cooperation among RRHs and an effective resource optimization.

In addition to the soon-to-come traffic surge from the mobile devices, it is expected that the future Internet will attract a large number of networked objects [6]. Such a network paradigm is referred to as IoT where the physical devices in the surrounding areas are connected to the Internet to collect data and/or to react to the events happening nearby. The IoT paradigm has begun to change the major portion of the economy, and before long, trillions of “Things” are expected to be connected to each other, and eventually, to the Internet [7]. The IoT has been bringing tremendous opportunities in many applications such as environment monitoring, smart home/city, automated manufacturing, ambient assisted living, inventory monitoring, smart grid, health, public safety and so forth.

One major technical challenge in realizing such IoT applications is the support of large-scale networked devices in a unified manner. So far, different applications have been using different networking systems to connect devices to achieve the application-specific goals; for example, Zigbee, IEEE 802.15.4, 6LoWPAN, Bluetooth and cellular networks, to name a few. Such diversification in networking solutions makes the cooperation among different networks challenging, and thus, limiting the scalability of the network. Another challenging aspect is the use of battery-operated devices. IoT devices in general are resource-constrained in that they have limited computing power, form factors and battery capacity. Considering that the expected lifespan of IoT applications are ranging from months to years, an energy-efficient operation of IoT devices is important to prolong the lifetime of the IoT devices and the IoT application.

To support such large-scale network of resource-constrained devices, IEEE released a new WLAN (wireless local area network) standard for IoT in 2017, called IEEE 802.11ah [8] [9] or Wi-Fi HaLow. IEEE 802.11ah provides a long transmission range (i.e., approximately 1km) and a support of a large number of devices to be associated (i.e., approximately 8000 devices per access point), and thus, is considered as one of the most promising networking solutions for large-scale IoT applications. Also, IEEE 802.11ah has a set of features to enhance the energy efficiency of battery-operated devices, for example, Power Saving Mode and Restricted Access Window mechanism by which many devices can stay in the low-power state for long.

1.2 Motivation

On one hand, experts from both academia and industry are seeking for ways to boost up the network capacity (e.g., throughput) to be prepared for the high traffic increase from bandwidth-hungry applications and large-scale IoT networks. On the other hand, the increased energy consumption is a matter of concern to the environment because of the greenhouse gas emission. It has been estimated that the Information and Communication Technology (ICT) sector accounts for 2% of the total CO_2 of emission. Among those from networks, to be specific, the mobile communication infrastructures will contribute more than 50% by 2020 [10]. In particular, the power consumption from BSs is expected to account for 60-80% of the total power usage from cellular networks [11] due to the rapid growth in both the number of BSs deployed and the aggregate mobile traffic.

The increased power consumption is also becoming one of the main sources for the high operational expenditure. In order to meet the ever-growing traffic demand from users, mobile network operators have been expanding the network infrastructure to increase the network capacity. Such a practice also increases the energy consumption from the infrastructure because, in general, BSs always need to be active to provide coverage and service availability. Since the majority of energy consumption is stemming from the BSs [12], an energy efficient resource management scheme for BSs needs to be thoroughly studied to reduce both the negative effect on the environment and the operational expenditure to the network operators.

The energy consumption has an importance in IoT applications as well, but in a different perspective. The major concern in the network of resource-constrained devices is that the IoT devices are battery-operated. To guarantee the expected longevity of IoT applications, IEEE 802.11ah has introduced a set of built-in power saving mechanisms including Power Saving Method. In addition, the structured device management scheme and the restricted access window method limit the the channel contention with which devices can spend less power on accessing channel and stay in the low-power state for long. However, further optimizing the energy-efficiency of battery-operated devices is essential to meet the expected lifetime of both the IoT devices and the network. It is particularly important to some critical applications such as battlefield, public safety and disaster monitoring, to protect citizens and their property.

1.3 Overall Objective

In this dissertation, we study how to optimally schedule the networking resources for better energy-efficiency and quality service. To do so, we first study to identify the essential resource components in the network of our interest, and then, propose a mathematical model/formulation representing how much resource to assign to each user/device to provide how much of the demand to satisfy. The energy efficiency which is the primary performance metric to be considered in this dissertation will be introduced as either a sole objective or part of the combined multiple objectives in each studies. In addition, each problem formulation to be introduced has a set of constraints to be satisfied, such as quality of service (QoS), minimal signal-to-interference-plus-noise ratio (SINR), and/or a set of rules to be abide by depending on the networking standard to operate with. The specific goal (or a set of goals) and the set of constraints will be clarified at the beginning section of each chapter.

1.4 Organization of Dissertation

The studies to be introduced in this dissertation can be grouped into the following two categories. One is the energy-efficient resource scheduling for cellular networks, and the other is the energy-efficient device scheduling for the resource-constrained IoT devices on the IEEE 802.11ah WLANs standard. For each study, we consider a certain networking standard or technology to be considered as one of the main drivers for the future networking system. In Chapter 2 and Chapter 3, we propose an energy-efficient resource scheduling scheme for the beyond-4G heterogeneous cellular networks and the 5G C-RANs, respectively, to provide a quality service to the cellular service subscribers. The Chapter 4 and Chapter 5, on the other hand, introduce an energy-efficient device scheduling and management algorithm, respectively, for the IEEE 802.11ah WLAN which is designed to provide a large-scale connectivity to the resource-limited and battery-operated IoT devices. To conclude, in Chapter 6, we summarize the dissertation, and then, introduce future research directions.

CHAPTER 2. QOS-AWARE ENERGY-EFFICIENT RESOURCE OPTIMIZATION FOR BEYOND-4G CELLULAR NETWORKS

2.1 Summary

A heterogeneous network (HetNet) can actively utilize the spectrum reuse with low power consumption, and thus is promising for the next-generation cellular networks. However, there are some technical challenges to be overcome in order for HetNets to be practical, and we address the following two in this chapter. One is how to formulate the association and resource scheduling problem in a way that an optimal solution can be found in a reasonable amount of time, and the other is how to accommodate varying users' demand. In order to minimize the power consumption and to satisfy varying users' QoS (Quality of service) requirements, we propose a low-complex, distributed association and resource allocation scheme. By taking a cost-based approach, we first separate a non-convex joint association and resource allocation problem into two subproblems. The channel allocation and base station assignment problem is then relaxed so that the problem becomes tractable. For the power allocation problem, we introduce a low-complex iterative algorithm by using the decomposition theory. The evaluation results show that the proposed solution can maintain the overall power consumption minimized while satisfying the QoS requirements. The complete version of this chapter has been published in [13].

2.2 Introduction

A recent report pointed out that in 2014 the amount of data traffic from mobile devices increased 69%, resulting in a monthly usage of 2.5 exabytes [4]. Such an explosive increase in mobile traffic has been led by a widespread usage of mobile handheld devices and bandwidth-hungry applications. The ever-increasing mobile traffic is unlikely to be saturated at least for the next five years; rather, the total amount of traffic generated by mobile devices is expected to be increased by tenfold between 2014 and 2019 [4].

On one hand, experts from academia and industry are seeking for ways of boosting up the communication technology to be prepared for the sharply-increasing traffic demand. On the other hand, the increased energy consumption is a matter of concern to the environment because of the greenhouse gas emissions and the increasing OPEX[1]. It is estimated that the ICT sector accounts for 2% of the total CO_2 emission, and among those from networks, to be specific, the mobile communication infrastructures will contribute more than 50% by 2020 [10]. In particular, the power consumption from BSs is expected to account for 60–80% of the total power usage from cellular networks [11] due to the rapid growth in both the number of BSs deployed and the aggregate mobile traffic.

Among those technologies taking the two aforementioned aspects into account, the heterogeneous architecture [14] is one of the most promising technologies, because it is not only applicable to the current 4G networking system, but also considered as an essential component for the future generation (5G) networking system [3][15][16]. A HetNet, in general, is formed by deploying multiple low-power, low-cost SBS (e.g., microcells, picocells and femtocells) on top of a high-power MBS, and it has many advantages. For example, due to the short coverage of SBSs, the spectrum reuse can be actively exercised. Also the channel quality between an SBS and its associated UE is so good that a higher data rate can be easily achieved, while operating in low power. Being relatively compact in size as well as the ease of installation enables SBSs to be flexibly deployed so that they can effectively extend the coverage of MBSs and offload users' traffic especially in crowded areas, such as shopping malls, sport stadiums, concert arenas and so forth. Further, much cheaper CAPEX and OPEX of SBSs have intrigued a great amount of attention from both academia and industry.

Motivated by the fact that HetNet is a cost-effective and practical solution, the use of SBSs has been widely introduced (including [3], [15] and [17]) and studied in many literatures. In addition,

[1] Abbreviations and acronyms are listed in Table 2.1.

Table 2.1: Abbreviations and acronyms used in Chapter 2

CAPEX	CAPital EXpenditure
HetNet	Heterogeneous Network
ICT	Information and Communications Technology
LTE	Long-Term Evolution
MBS	Macro Base Station
OFDMA	Orthogonal Frequency-Division Multiple Access
OPEX	OPerational EXpenditure
QoS	Quality of Service
SBS	Smallcell Base Station
SINR	Signal-to-Interference-plus-Noise Ratio
UE	User Equipment

the concept of HetNet is accepted by the standard body and introduced in LTE [14] [18] [19]. One common concern in those works is either how to offload the user traffic from an MBS to SBSs or how to schedule the network resource so as to achieve their own objectives, such as maximizing spectral efficiency, energy efficiency, a certain utility measure and so forth.

As mentioned in [20], however, introducing SBSs may rather increase the overall power consumption unless they are handled carefully. In general, the careful handling includes well-designed offloading strategies (i.e., establishing associations between UEs and SBSs), preferably with a dynamic switching on/off scheme [10][11], and resource scheduling (i.e., transmission power control and bandwidth allocation). Even after the optimal decision has been made, one cannot guarantee the optimal decision will remain optimal in the future because the QoS requirement of each user and/or the wireless link quality frequently changes over time. However, a frequent manual reconfiguration of the system is not appealing because it is hard to be responsive to the network dynamics, and also increases OPEX to a large extent especially when the BSs are densely deployed[19]. In this regard, a self-configuring or an automated mechanism that is able to respond or adapt to the network dynamics needs to be studied from the perspective of HetNets.

In addition, the centrality can cause a serious issue especially in HetNets. In a centralized system, all decisions are made by one or a small group of entities. As a result, a huge volume of information should be either exchanged in real-time or stored/updated at a shared storage, which incurs a significant burden to the system. In addition, a centralized system requires a high processing capability and power consumption to handle a large volume of data as the network grows in size. For those reasons, a centralized system may not be responsive and it might even consume a large amount of power for resource scheduling; whereas a distributed algorithm does not.

In this chapter, we study the distributed user association and resource allocation for HetNets that minimizes the overall power consumption. By considering both association and resource scheduling together, we propose a complete management framework for an energy-efficient HetNet. In order to efficiently schedule the use of BSs as well as the networking resource (i.e., transmission power and spectrum) we formulate a two-stage iterative optimization problem which can be solved in a distributed manner without requiring a heavy control message exchange or a high computational cost. To do so, we first partition the non-convex, joint association and resource allocation problem into two subproblems by using a cost-based approach that effectively estimates the power use. In addition, relaxation and decomposition techniques are applied to the association and the resource scheduling problems, respectively, so as to reduce the computational complexity. After the decomposition, we propose a distributed power update method, which converges to the optimum; we show its convergence by both simulation and analysis. An extensive amount of evaluations and comparison studies has been performed on various network scenarios to show the effectiveness of the proposed BS/channel assignment and power allocation scheme. Lastly, we show the efficiency of the proposed scheme by showing its fast convergence.

The rest of this chapter is organized as follows. In Section 2.3, we describe the network model and then introduce the problem formulation for both i) BS association and channel assignment and ii) power allocation. Section 2.4 presents the evaluation and comparison results, and finally Section 2.5 concludes the chapter.

2.3 Problem Formulation

We begin this section by describing the network model and assumptions. In what follows, we introduce how the optimal user association and resource allocation problem are formulated.

2.3.1 Network Model and Assumptions

Throughout the chapter, we focus on the downlink transmission for a two-tier OFDMA (Orthogonal Frequency-Division Multiple Access) cellular network. On the network are M MBSs, and each of which is overlaid by S SBSs. MBSs are distributed in a planned manner (e.g., by keeping the same inter-cell distance between the nearby MBSs) in order to provide area coverage, mobility management and so forth; while SBSs are randomly[2] distributed [14] [23] [26] [27] [40] [41] [42]. The reason for assuming the random deployment of SBSs is that their deployment is much less planned [20] compared to that of MBSs. To be specific, they are likely to be installed on demand or in an ad hoc manner so as to fulfill a sudden or periodic increase of the QoS requirement in certain areas such as a shopping center, sports complex, office, household and so forth. It is worth mentioning that the proposed scheme does not assume or rely on any specific area or region. Therefore, in order to show that the proposed scheme does not depend on any certain distributions of SBS, a random distribution is used to represent the placement/layout of SBSs in general. All MBSs are always active in order to provide the full area coverage to all UEs [19]. On the other hand, SBSs may or may not be active depending on the user association status at the moment. Any non-offloaded UEs are automatically associated with the MBS that provides the strongest signal strength by default.

MBSs and SBSs are assumed to operate on different frequency bands to avoid cross-tier interference [31] [32] [34] [42]. However, the same type of base stations share the same frequency range, and thus they always interfere with each other. In this work, the *coverage* of an MBS indicates the area within which all non-offloaded UEs shall associate with the MBS. However, the signal generated by each MBS propagates beyond its coverage, and thus incurs interference to the rest MBSs; this principle applies to SBSs as well. Both types of base stations can access the core network through wired communication links. Each MBS has U UEs (or users) whose average data rate requirements are known. We assume that SBSs operate fully (or in part) with an open access mode.[3] The available bandwidth is divided into multiple channels, each of which is Δf-wide in Hz.

[2] In Section 2.4, we have used Uniform distribution for simulation.

[3] If all the SBSs are deployed by the end-users, the open access mode may not be a practical assumption to make. However, by focusing on the scenario where all SBSs are deployed by the network operator, or considering only those SBSs operating with the open access mode (or the hybrid mode [20]), we argue that this assumption still holds.

Table 2.2: Summary of notations used in Chapter 2

M	Number of MBSs on the network
S	Number of SBSs on a macrocell
U	Number of UEs on a macrocell
N_{ch}	Number of available channels
N_{ch}^{M}	Number of available channels for an MBS
N_{ch}^{S}	Number of available channels for an SBS
$\mathcal{N}^{M}$	Index set of MBS-accessible channels
$\mathcal{N}^{S}$	Index set of SBS-accessible channels
$\mathcal{M}$	Index set of MBSs, $\{1, 2, \cdots, m, \cdots, M\}$
$\mathcal{S}^{m}$	Index set of SBSs overlaid on MBS m, $\{1, 2, ..., s, ..., S\}$
$\mathcal{U}^{m}$	Index set of UEs within the coverage of MBS m, $\{1, 2, ..., u, ..., U\}$
$\mathcal{U}_0^{m}$	Subset of $\mathcal{U}^m$. UEs associated with MBS m
$\mathcal{U}_s^{m}$	Subset of $\mathcal{U}^m$. UEs associated with SBS $s \in \mathcal{S}^m$
$\mathbf{p}_0^{m}$	Transmission power vector of MBS m over channels
$\mathbf{p}_s^{m}$	Transmission power vector of SBS $s \in \mathcal{S}^m$ over channels
$\mathbf{g}_u^{m}$	Channel gain vector of UE $u \in \mathcal{U}^m$ over channels
η_{thr}	SINR threshold
r_u^{m}	QoS (i.e., data rate) requirement of UE $u \in \mathcal{U}^m$
Δf	Channel bandwidth
P_{max}^{M}	Maximum transmission power of MBS
P_{max}^{S}	Maximum transmission power of SBS
$\mathbf{c}_u^{m}$	Cost vector of UE $u \in \mathcal{U}^m$ over channels
N_0	Per-Hz noise power

We also assume the continuous power and rate control. Some of the frequently-used notations are summarized in Table 2.2, and other notations that are not on the table will be introduced when necessary.

2.3.2 Cost-Based Problem Separation

Considering that the joint association and resource allocation problem belongs to a mixed integer nonlinear program with the decision variables coupled, the computational cost of the joint problem is prohibitive. Thus, the approach taken in this work is to first partition the problem into two, one for the user association and channel allocation (Stage 1), and the other for the power allocation (Stage 2), and then apply relaxation and decomposition techniques, respectively, in order to make the whole procedure tractable and suitable for online resource scheduling. To this end, we have introduced a cost function, by which the original problem will be separated into two. It is worth mentioning that after the partitioning, the proposed two-staged method may result in a sub-optimal solution. After the problem separation, however, Stage 1 does not know how much power will actually be used for communication. Therefore, it is crucial that the cost needs to be designed in a way that it can correctly estimate the amount of power to be used.

At the beginning of Stage 1, a UE $u \in \mathcal{U}^m$ senses the pilot signals from nearby BSs over the entire channels, and produces a channel gain vector $\mathbf{g}_u^m \in \mathbb{R}_+^{N_{ch}}$. Out of N_{ch} entries in $\mathbf{g}_u^m$, N_{ch}^M elements correspond to the measured channel gains between UE $u \in \mathcal{U}^m$ and MBS m over N_{ch}^M channels that are accessible to MBSs. Therefore, all the N_{ch}^M elements in $\mathbf{g}_u^m$ are strictly positive due to the area coverage provided by the closest MBS m, while the rest elements are nonnegative. If a UE resides within the coverage of an SBS, we have $\mathbf{g}_u^m \succ \mathbf{0}_{N_{ch}}$ where $\succ$ is an element-wise greater than operator and $\mathbf{0}_{N_{ch}}$ is an N_{ch}-by-1 zero vector. Also, in such cases, a UE can recognize the identifier of the SBS by decoding the pilot signal.

To be consistent throughout the chapter, let us assume that the indices of the channels used by MBSs come before the ones used by SBSs. In other words, out of N_{ch} number of channels available whose index starts from 1 to N_{ch}, each MBS has an access to the channels indexed by $1, 2, \cdots, N_{ch}^M$, while an SBS is allowed to use the ones indexed by $N_{ch}^M + 1, \cdots, N_{ch}$. In this regard, let $\mathcal{N}^M$ and $\mathcal{N}^S$ be $\{1, 2, \cdots, n, \cdots, N_{ch}^M\}$ and $\{N_{ch}^M + 1, \cdots, n, \cdots, N_{ch}\}$, respectively.

Given $\mathbf{g}_u^m$ and the data rate requirement $r_u^m \in \mathbb{R}_{++}$, each UE u builds its own cost vector $\mathbf{c}_u^m \in \mathbb{R}_{++}^{N_{ch}}$, where its n-th entry is:

$$c_{u,n}^m = \begin{cases} (2^{\tilde{r}_u^m} - 1)/g_{u,n}^m, & \text{if } g_{u,n}^m > 0. \\ \infty, & \text{otherwise,} \end{cases} \tag{2.1}$$

where $\tilde{r}_u^m = r_u^m/\Delta f$ is a normalized data rate requirement. In fact, $\mathbf{c}_u^m$ is a measure of power required to satisfy the data rate requirement of UE $u \in \mathcal{U}^m$ across all channels with the interference and noise term ignored. According to the Shannon's well-known channel capacity formula, an achievable bit rate over a channel is defined as $C = B\log_2(1+\frac{gP}{I+N})$, where C is channel capacity measured in bits per second (bps), g is channel gain, P is transmission power, I is interference and N is noise. In order not to violate the QoS requirement for each UE, we need to satisfy the constraint $r_n^m \geq C$. Since the minimum power is achieved when the QoS requirement is satisfied with equality, we can rewrite the Shannon's capacity formula as follows after replacing B and g with Δf and $g_{u,n}^m$, respectively, to be consistent in notation: $r_u^m = \Delta f \cdot \log_2(1+\frac{g_{u,n}^m P}{I+N})$ or $\tilde{r}_u^m = \frac{r_u^m}{\Delta f} = \log_2(1+\frac{g_{u,n}^m P}{I+N})$. After rearranging the equation, the minimum transmission power that satisfies the QoS requirement can be found by $P = (2^{\tilde{r}_u^m})(I+N)/g_{u,n}^m$. By assuming the sum of interference and noise is not dominant in determining the transmission power, we get the following relation which leads to how we defined the cost term in Eq. (2.1), i.e., $P \propto 2^{\tilde{r}_u^m}/g_{u,n}^m = c_{u,n}^m$.

After gathering the cost vectors from all UE ($\forall u \in \mathcal{U}^m$) along with their nearby SBS IDs, if applicable, an MBS m runs both Stage 1 and 2 in sequence which will be discussed as follows.

2.3.3 Stage 1: User Association and Channel Assignment

The goal of this stage is to find the best association (i.e., an offloading strategy) and channel assignment that minimizes the overall cost, which represents the expected amount of power use as discussed before. Given the cost vectors collected, we have the following optimization problem P. 2.2 for each MBS m that minimizes the overall cost of making user association and channel assignment. Please note that *s.t.* in the problem formulation stands for *subject to.*

$$\min_{\mathbf{X}^m} \quad tr[\mathbf{X}^m \cdot (\mathbf{c}^m)^T] \tag{2.2a}$$

$$\text{s.t.} \quad \sum_{n=1}^{N_{ch}} X_{u,n}^m = 1, \forall u \in \mathcal{U}^m \tag{2.2b}$$

$$\sum_{u \in \mathcal{U}^m} X_{u,n}^m \leq 1, \forall n \in \mathcal{N}^M \tag{2.2c}$$

$$\sum_{u \in \mathcal{U}^m} X_{u,n}^m \cdot I_{u,s}^m \leq 1, \forall n \in \mathcal{N}^S, \forall s \in \mathcal{S}^m \tag{2.2d}$$

$$\mathbf{X}^m \in \{0,1\}^{U \times N_{ch}}, \tag{2.2e}$$

where $tr[\cdot]$ is the trace function that sums the diagonal elements of a matrix $\cdot$, the decision variable $\mathbf{X}^m$ is a U-by-N_{ch} matrix of which (u, n) element is 1 (or 0) if UE u is (or is not) assigned to channel n, $\mathbf{c}^m$ is a U-by-N_{ch} matrix whose u-th row corresponds to the cost vector of UE u and $\mathbf{I}^m$ is an U-by-S matrix whose (u, s) element is 1 if UE u successfully decoded the pilot signal from SBS s. The objective function (2.2a) in P. 2.2 calculates the total cost with respect to the mapping between UEs and channels (and BS as well). The objective function can also be written as $\sum_{\forall u \in \mathcal{U}^m} \mathbf{X}_u^m \cdot (\mathbf{c}_u^m)^T$ or $\sum_{\forall u \in \mathcal{U}^m} \sum_{n=1}^{N_{ch}} X_{u,n}^m \cdot c_{u,n}^m$. Since the cost $X_{u,n}^m \cdot c_{u',n'}^m$ is meaningful only when $u = u'$ and $n = n'$, we take the sum of only the diagonal elements from $\mathbf{X}^m \cdot (\mathbf{c}^m)^T$, i.e., $tr[\mathbf{X}^m \cdot (\mathbf{c}^m)^T]$. Each UE is allowed to use 1 unit of channel resource (2.2b), and each MBS and SBS channel cannot be used for more than 1 unit each, (2.2c) and (2.2d), respectively. The decision variable represents a membership relation, and thus is binary (2.2e).

Please note that P. 2.2 runs on a small time scale; for example, 1 ms to comply with the 3GPP E-UTRA requirement [18]. Given that 3GPP E-UTRA makes use of physical resource blocks for communication, even when the number of available channels is less than that of active UEs, the proposed method can still fulfill the service demand from UEs by scheduling the resource blocks. As long as the demand from a UE can be satisfied without violating the delay constraint, the proposed method may schedule the UE for communication in one of the following time slots if the number of available channels at the moment is not enough. That is, having failed in assigning a BS/channel to a UE for the moment does not necessarily mean a failure in satisfying the UE's QoS requirement. In addition, the densely deployed, short-range SBSs can achieve high frequency reuse, meaning that the aggregate number of channels seen by users can be larger than that of physical channels. However, if the aggregate service demand for a certain period exceeds the maximum attainable throughput over the network during the same period, some of the active users may experience service degradation, which will be discussed in Section 2.3.4.

Due to the combinatorial nature of P. 2.2, however, the problem is not tractable, and thus is not suitable for an online scheduling. By relaxing the binary constraint (2.2e), we get the following convex problem that can be efficiently solved by each MBS m with the complexity of $\mathcal{O}((U \cdot N_{ch})^3)$

when the interior point method is used.

$$
\begin{aligned}
\min_{\mathbf{X}^m} \quad & tr[\mathbf{X}^m \cdot (\mathbf{c}^m)^T] && (2.3a)\\
\text{s.t.} \quad & \sum_{n=1}^{N_{ch}} X_{u,n}^m = 1, \forall u \in \mathcal{U}^m && (2.3b)\\
& \sum_{u \in \mathcal{U}^m} X_{u,n}^m \leq 1, \forall n \in \mathcal{N}^M && (2.3c)\\
& \sum_{u \in \mathcal{U}^m} X_{u,n}^m \cdot I_{u,s}^m \leq 1, \forall n \in \mathcal{N}^S, \forall s \in \mathcal{S}^m && (2.3d)\\
& \mathbf{X}^m \in [0,1]^{U \times N_{ch}}. && (2.3e)
\end{aligned}
$$

Although the optimal solutions from both P. 2.2 and P. 2.3 indicate the channel assignment as well as the user association for each UE, both solutions are not the same in practice. The binary solution from P. 2.2 lets each UE use the assigned channel and BS for a unit time, whereas the (possibly) non-binary solution from P. 2.3 forces a UE to hop between channels (and possibly between BSs as well) during the same unit time since the optimal solution indicates the fraction of time that a UE is allowed to use one or more BSs and channels. The non-binary solution seems to be attractive since it yields a better (or at least the same) optimal value because of the relaxation. However, it increases the amount of control messages as well as the scheduling complexity due to the frequent handover and channel hopping that should be made on a very precise timescale.

In this regard, we will recover binary solutions from the non-binary solutions from the relaxed optimization problem P. 2.3 by using the one-by-one removal algorithm [36] [43] [44] [45]. This relax-and-recover approach will help the system maintain a low complexity in both computation and operation.

The Algo. 1 iteratively solves the relaxed optimization problem (line 2), searches for the nonzero minimum value for each UE (line 4), and forces each to be zero (line 5). Each MBS concurrently runs the algorithm which terminates in less than or equal to N_{ch} number of iterations. Please note that the solution found by running Algo. 1 may yield a sub-optimal solution to P. 2.2; the optimality of the solution will be investigated in Section 2.4. Given the recovered binary user association and channel assignment decision, the following Stage 2 allocates the minimum power to each UE. Note

Algorithm 1 Gradual one-by-one removal (for MBS m)

```
repeat
    Solve the relaxed optimization problem P. 2.3
    for $\forall u \in \mathcal{U}^m$ do
        $n^* = \arg\min_{\forall n} X_{u,n}^m$ such that $X_{u,n}^m \neq 0$
        Set $X_{u,n^*}^m = 0$
    end for
until all $X_{u,n}^m$ are binary
```

that the channel assignment problem also finds the best BS match for each UE. Those SBSs with no UE associated shall change their state into the SLEEP mode in order to save energy, while the other SBSs stay in the ACTIVE mode [20].

2.3.4 Stage 2: Power Allocation

This stage allocates the minimum power to each UE by taking both SINR and QoS requirements into account. In this regard, we first formulate the centralized power allocation problem where one or a small number of central entities have to control the downlink power for all BSs. In what follows, the centralized problem is decomposed into multiple low-complex subproblems which are scalable and suitable for online scheduling.

Before introducing the Stage 2 problem formulation, let us extend the notation of the channel gain so that we can comprehensively represent the gain between all the entities including that do not even belong to the same macrocell. As a reminder, the channel gain $\mathbf{g}_u^m$ represents the channel gain between UE $u \in \mathcal{U}^m$ and either MBS m or SBS $s \in \mathcal{S}^m$, where all of them are within the coverage of MBS m. This is because $\mathbf{g}_u^m$ is determined by overhearing the pilot signals over channels; thus, it should be coupled with the nearest MBS and SBS (if applicable). Let $G_{u,n}^{0,m}$ be the gain over channel n between UE u and MBS m, where u does not need to be a member of $\mathcal{U}^m$. In the same manner, let $G_{u,n}^{s,m}$ be the gain over channel n between UE u and SBS s, where u does not need to be a member of $\mathcal{U}^m$, but s must be a member of MBS m, i.e., $s \in \mathcal{S}^m$. Thus, for any

UE $u \in \mathcal{U}^m$, we have $\mathbf{g}^m_{u,n} = G^{0,m}_{u,n}$ for any $n \in \mathcal{N}^M$. On the other hand, we have $\mathbf{g}^m_{u,n} \neq G^{0,m'}_{u,n}$ for any $n \in \mathcal{N}^M \bigcup \mathcal{N}^S$ if $m \neq m'$.

Given the BS association and channel assignment made in Stage 1, we have the power allocation problem P. 2.4 for MBS m and all SBSs therein (i.e., $\forall s \in \mathcal{S}^m$) that minimizes the overall power consumption.

$$\min_{\mathbf{P}^m} \sum_{n \in \mathcal{N}^M} P^m_{0,n} + \sum_{s \in \mathcal{S}^m} \sum_{n \in \mathcal{N}^S} P^m_{s,n} \tag{2.4a}$$

s.t.

$$\eta^m_{u,n} \geq \eta_{thr} X^m_{u,n}, \forall n \in \mathcal{N}^M, \forall u \in \mathcal{U}^m_0 \tag{2.4b}$$

$$\eta^m_{u,n} \geq \eta_{thr} X^m_{u,n}, \forall n \in \mathcal{N}^S, \forall u \in \mathcal{U}^m_s, \forall s \in \mathcal{S}^m \tag{2.4c}$$

$$\Delta f \log_2(1 + \eta^m_{u,n}) \geq r^m_u X^m_{u,n}, \forall n \in \mathcal{N}^M, \forall u \in \mathcal{U}^m_0 \tag{2.4d}$$

$$\Delta f \log_2(1 + \eta^m_{u,n}) \geq r^m_u X^m_{u,n}, \qquad \forall n \in \mathcal{N}^S, \forall u \in \mathcal{U}^m_s, \forall s \in \mathcal{S}^m \tag{2.4e}$$

$$0 \leq P^m_{0,n} \leq P^M_{max}, \forall n \in \mathcal{N}^M \tag{2.4f}$$

$$\sum_{n \in \mathcal{N}^M} P^m_{0,n} \leq P^M_{max} \tag{2.4g}$$

$$0 \leq P^m_{s,n} \leq P^S_{max}, \forall n \in \mathcal{N}^S, \forall s \in \mathcal{S}^m \tag{2.4h}$$

$$\sum_{n \in \mathcal{N}^S} P^m_{s,n} \leq P^S_{max}, \forall s \in \mathcal{S}^m, \tag{2.4i}$$

where $P^m_{0,n}$ is the power allocated by MBS m over channel $n \in \mathcal{N}^M$, $P^m_{s,n}$ is the power allocated by SBS $s \in \mathcal{S}^m$ over channel $n \in \mathcal{N}^S$, and $\eta^m_{u,n}$ in Eq. (2.4b) and Eq. (2.4c) is the measure of SINR defined in Eq. (2.5a) if $u \in \mathcal{U}^m_0$ and Eq. (2.5b) if $u \in \mathcal{U}^m_s$, respectively. If a UE $u \in \mathcal{U}^m$ is to associate with an MBS m (i.e., $u \in \mathcal{U}^m_0$) on a certain channel $n \in \mathcal{N}^M$, the interference that the UE u will experience is related to the transmission power allocated to the same channel by the other MBSs $m' \neq m$, which corresponds to Eq. (2.5a). On the other hand, if a UE is coupled with an SBS on a certain channel $n \in \mathcal{N}^S$, it will sense the interference caused by all the other SBSs on the network that have allocated transmission power to the same channel as in Eq. (2.5b). To

$$\eta_{u,n}^{m} = \begin{cases} \dfrac{G_{u,n}^{0,m} P_{0,n}^{m}}{\sum_{m' \neq m \in \mathcal{M}} G_{u,n}^{0,m'} P_{0,n}^{m'} + \Delta f \cdot N_0} & \text{(2.5a)} \\ \dfrac{G_{u,n}^{s,m} P_{s,n}^{m}}{\sum_{s' \neq s \in \mathcal{S}^m} G_{u,n}^{s',m} P_{s',n}^{m} + \sum_{m' \neq m \in \mathcal{M}} \sum_{s' \in \mathcal{S}^{m'}} G_{u,n}^{s',m'} P_{s',n}^{m'} + \Delta f \cdot N_0} & \text{(2.5b)} \end{cases}$$

be specific, in Eq. (2.5b) the first term in the denominator measures the interference from other SBSs in the same macrocell, whereas the second term measures the interference from all SBSs that do not belong to the same macrocell. Note that the amount of interference to an SBS-associated UE is not significant mainly due to the low transmission power of SBSs, and the penetration loss of walls. For each UE that is associated with an MBS or SBS, its SINR should be greater than or equal to the predefined threshold, η_{thr}, as in Eq. (2.4b) and Eq. (2.4c), respectively. The QoS requirements of UEs that are associated with an MBS or an SBS should be satisfied according to Eq. (2.4d) or Eq. (2.4e), respectively. The transmission power allocated to a certain channel cannot exceed the power budget of an MBS or an SBS as in Eq. (2.4f) or Eq. (2.4h), respectively. Finally, the aggregate transmission power of an MBS or an SBS cannot be larger than its power budget as denoted by Eq. (2.4g) or Eq. (2.4i), respectively.

It is worth mentioning that solving P. 2.4 for MBS m and all active SBSs therein is not independent of that of others since each MBS m and all SBSs therein need to know the inter-tier interference from the rest MBSs and SBSs, respectively. Therefore, a single or a set of computing resource has to solve the network-wide power allocation problem, making the centralized approach impractical for an online resource scheduling. In what follows, we transform the centralized power allocation problem P. 2.4 into low-complex subproblems such that each subproblem can be quickly solved in a distributed manner.

In order to build a distributed system we decompose the centralized problem P. 2.4 by using the decomposition theory [46], and then transform it into low-complex subproblems that can be independently solved by each BS. The power allocation problem P. 2.4 which is for both MBS m and all SBSs therein (i.e., $s \in \mathcal{S}^m$) already has two sets of easily-separable components. The first term in the objective function (2.4a) along with the following four constraints (2.4b), (2.4d), (2.4f) and (2.4g) forms the MBS power minimization problem which is independent of that for SBSs, i.e., the remaining parts of the problem. Therefore, we can form the power allocation problem only for MBS m as follows.

$$\min_{\mathbf{P}_0^m} \sum_{n\in\mathcal{N}^M} P_{0,n}^m \tag{2.6a}$$

s.t.

$$\eta_{u,n}^m \geq \eta_{thr} X_{u,n}^m, \forall n \in \mathcal{N}^M, \forall u \in \mathcal{U}_0^m \tag{2.6b}$$

$$\Delta f \log_2(1+\eta_{u,n}^m) \geq r_u^m X_{u,n}^m, \forall n \in \mathcal{N}^M, \forall u \in \mathcal{U}_0^m \tag{2.6c}$$

$$0 \leq P_{0,n}^m \leq P_{max}^M, \forall n \in \mathcal{N}^M \tag{2.6d}$$

$$\sum_{n\in\mathcal{N}^M} P_{0,n}^m \leq P_{max}^M. \tag{2.6e}$$

What is left in P. 2.4 after taking P. 2.6 out is the power minimization problem for all SBSs $s \in \mathcal{S}^m$, where its objective is to minimize the sum of transmission power used by all SBSs in macrocell m with the following constraints, (2.4c), (2.4e), (2.4h) and (2.4i). Minimizing the total power usage is equivalent to minimizing each individually. Also, the set of constraints for each SBS s is independent of that for the rest SBSs provided the transmission power of other SBSs are fixed. As a result, we have the power allocation problem for each SBS $s \in \mathcal{S}^m$ as follows which can be solved if the transmission power and the channel gain information of other SBSs are assumed to be known.[4]

$$\min_{\mathbf{P}_s^m} \sum_{n\in\mathcal{N}^S} P_{s,n}^m \tag{2.7a}$$

s.t.

$$\eta_{u,n}^m \geq \eta_{thr} X_{u,n}^m, \forall n \in \mathcal{N}^S, \forall u \in \mathcal{U}_s^m \tag{2.7b}$$

$$\Delta f \log_2(1+\eta_{u,n}^m) \geq r_u^m X_{u,n}^m, \forall n \in \mathcal{N}^S, \forall u \in \mathcal{U}_s^m \tag{2.7c}$$

[4] Please note that the proposed method does not directly solve P. 2.7 and thus, it does not really make such assumptions. In fact, P. 2.7 is one of the steps that we make to design the distributed power allocation method which will yield the global optimal solution without making the assumptions.

$$0 \leq P^m_{s,n} \leq P^S_{max}, \forall n \in \mathcal{N}^S \tag{2.7d}$$

$$\sum_{n \in \mathcal{N}^S} P^m_{s,n} \leq P^S_{max}. \tag{2.7e}$$

Although both P. 2.6 and P. 2.7 as they are cannot be further decomposed due to the coupling constraints in (2.6e) and (2.7e), respectively, we can use the decomposability structure by forming a Lagrangian of each to make both problems be decomposed. By relaxing (2.6e), the Lagrangian of P. 2.6 is given as below.

$$\begin{aligned} \min_{\mathbf{P}^m_0} \quad & \sum_{n \in \mathcal{N}^M} P^m_{0,n} + \lambda\Big(\sum_{n \in \mathcal{N}^M} P^m_{0,n} - P^M_{max}\Big) \\ \text{s.t.} \quad & \text{constraints in: } (2.6b), (2.6c), (2.6d), \end{aligned} \tag{2.8a}$$

where λ is a nonnegative Lagrangian multiplier. As a result, we have a Lagrange dual problem as follows.

$$\begin{aligned} \max_{\lambda \geq 0} \min_{\mathbf{P}^m_0} \quad & \sum_{n \in \mathcal{N}^M} P^m_{0,n} + \lambda\Big(\sum_{n \in \mathcal{N}^M} P^m_{0,n} - P^M_{max}\Big) \\ \text{s.t.} \quad & \text{constraints in: } (2.6b), (2.6c), (2.6d). \end{aligned} \tag{2.9a}$$

We assume that each BS has multiple processors or at least a single processor with the multi-threading capability, each of which is dedicated to each channel for power update. The dedicated processor or thread to each channel is called *channel manager*, which is in charge of controlling the downlink transmission power of the assigned channel. At the lower level, the channel manager for channel $n \in \mathcal{N}^M$ solves the following power minimization problem if there is a UE u associated with the channel (i.e., $X^m_{u,n} = 1$).

$$\min_{P^m_{0,n}} \quad h^m_{0,n}(\lambda) = (1+\lambda)P^m_{0,n} \tag{2.10a}$$

$$\text{s.t.} \quad \eta^m_{u,n} \geq \eta_{thr} \tag{2.10b}$$

$$\Delta f \log_2(1 + \eta^m_{u,n}) \geq r^m_u \tag{2.10c}$$

$$0 \leq P^m_{0,n} \leq P^M_{max}. \tag{2.10d}$$

Then, the higher level problem forms a maximization problem over the Lagrange multiplier as follows.

$$\max_{\lambda \geq 0} \quad h_0^m(\lambda) = \sum_{n \in \mathcal{N}^M} h_{0,n}^m(\lambda) - \lambda P_{max}^M \tag{2.11a}$$

Since the dual function $h_0^m(\lambda)$ is differentiable, the higher level problem can be solved with a gradient method of which update method is given below.

$$\lambda^{t+1} = [\lambda^t + \alpha^t(\sum_{n \in \mathcal{N}^M} P_{0,n}^{m*} - P_{max}^M)]^+, \tag{2.12}$$

where t is a nonnegative, integer-valued iteration count, α is a positive stepsize, and $[\cdot]^+ = \max\{0, \cdot\}$ is a projection operator to the nonnegative orthant. The initial λ can be set to some non-negative value, e.g., zero, and α can be a sufficiently small positive number; please refer to [46] for further details on the step-size. Then, the dual variable λ^t will converge to the dual optimal λ^* as $t \rightarrow \infty$ [46].

By taking a closer look at the lower-level problem P. 2.10, we can further simplify the power allocation procedure, and find a simple method to solve it by an even more efficient way than the decomposed ones. Since λ is nonnegative and common to all lower level problems, dropping the $1 + \lambda$ term from the objective function does not change the optimal decision value. For a power minimization problem with an SINR constraint, the optimality is achieved when the constraint is satisfied with equality, which is also true for the same problem with a QoS constraint. Therefore, the solution of P. 2.10 can simply be found by:

$$P_{0,n}^{m*} = \min\{\max\{P_{0,n}^{m(s)}, P_{0,n}^{m(q)}\}, P_{max}^M\}, \tag{2.13}$$

where $P_{0,n}^{m(s)}$ and $P_{0,n}^{m(q)}$ are the solutions that satisfy SINR and QoS requirements, respectively, with equality.

In the same manner, we can derive the distributed power allocation method for each SBS. By relaxing (2.7e) which is the coupling constraint in P. 2.7, we have the following Lagrangian.

$$\min_{\mathbf{P}_s^m} \sum_{n \in \mathcal{N}^S} P_{s,n}^m + \lambda(\sum_{n \in \mathcal{N}^S} P_{s,n}^m - P_{max}^S) \tag{2.14a}$$

$$\text{s.t.} \quad \text{constraints in: } (2.7b), (2.7c), (2.7d),$$

where λ is a nonnegative Lagrangian multiplier. At the lower level, the channel manager for channel $n \in \mathcal{N}^S$ solves the following power minimization problem if there is a UE associated with the channel (i.e., $X_{u,n}^m = 1$).

$$\min_{P_{s,n}^m} \quad h_{s,n}^m(\lambda) = (1+\lambda)P_{s,n}^m \tag{2.15a}$$

$$\text{s.t.} \quad \eta_{u,n}^m \geq \eta_{thr} \tag{2.15b}$$

$$\Delta f \log_2(1+\eta_{u,n}^m) \geq r_u^m \tag{2.15c}$$

$$0 \leq P_{s,n}^m \leq P_{max}^S. \tag{2.15d}$$

Considering the optimality condition of the given power minimization problem P. 2.15, its solution can be found by the following simple method as we did for each MBS channel manager.

$$P_{s,n}^{m*} = \min\{\max\{P_{s,n}^{m(s)}, P_{s,n}^{m(q)}\}, P_{max}^S\}, \tag{2.16}$$

where $P_{s,n}^{m(s)}$ and $P_{s,n}^{m(q)}$ are the solutions that satisfy SINR and QoS requirements, respectively, with equality.

Although each channel manager considers and guarantees the per-channel power budget constraint (i.e., the maximum transmission power for the channel should not be greater than P_{max}^M or P_{max}^S), it does not guarantee the per-BS power budget constraint (i.e., the total power use over all channels should not be greater than P_{max}^M or P_{max}^S) is also satisfied. For example, for $n, n+1 \in \mathcal{N}^M$, having $P_{0,n}^m = P_{max}^M$ and $P_{0,n+1}^m = P_{max}^M$ at the same time does not violate the power budget constraint of each channel manager. However, that is not a feasible solution because the sum transmission power over channels cannot exceed P_{max}^M. Therefore, the upper level entity should check whether the sum power constraint is violated or not. To this end, we use a simple policy for the upper level entity that if the aggregate power budget constraint is violated, let all active channel managers use the same transmission power.[5] To be specific, for each MBS m of which sum power constraint is violated, let each channel manager with an associated UE allocate

[5] Assigning the same power to all active channels provides close-to-optimal performance [47].

the transmission power in the following manner, $P_{0,n}^m = P_{max}^M / |\mathcal{U}_0^m|$, where $|\cdot|$ is the cardinality of a set $\cdot$. For each active SBS s of which sum power constraint is violated, let each channel manager with an associated UE use the power as $P_{s,n}^m = P_{max}^S / |\mathcal{U}_s^m|$. On the other hand, as long as the sum power constraint is satisfied, the upper level entity does not interrupt the power update procedures at lower-level channel managers.

2.3.5 Convergence

By using the analytical framework for convergence presented in [48], we prove that the proposed distributed power control algorithms in Eq. (2.13) and Eq. (2.16) converge to their corresponding optimum. Since both algorithms share the same structure and do not interfere with each other, we prove the convergence for an MBS m, i.e., Eq. (2.13). However, the following proof can be easily applied to the case for any SBS $s \in \mathcal{S}^m$, i.e., Eq. (2.16). To begin with, if the feasible power region is empty, i.e., if the power allocation problem P. 2.6 is infeasible, the transmission power for each channel converges (abruptly) to $P_{max}^M / |\mathcal{U}_0^m|$. This is because each channel manager is forced to use the equal transmission power when the sum power constraint is violated. In order to show convergence of the proposed scheme for the case of having a non-empty feasible power region, what follows is to transform the proposed power control method to the form of *interference function*, which is one of the key components in convergence analysis proposed in [48].

Since we consider only the case that the feasible power region is nonempty, we can simplify P. 2.6 by ignoring both constraints (2.6d) and (2.6e). After rearranging (2.6c), then, we get the following problem P. 2.17.

$$\min_{\mathbf{P}_0^m \succeq 0} \sum_{n \in \mathcal{N}^M} P_{0,n}^m \tag{2.17a}$$

$$\text{s.t.} \quad \eta_{u,n}^m \geq \eta_{thr} \cdot X_{u,n}^m, \forall n \in \mathcal{N}^M, \forall u \in \mathcal{U}_0^m \tag{2.17b}$$

$$\eta_{u,n}^m \geq 2^{\tilde{r}_u^m \cdot X_{u,n}^m} - 1, \forall n \in \mathcal{N}^M, \forall u \in \mathcal{U}_0^m. \tag{2.17c}$$

Since the optimality of P. 2.17 is achieved when among the two constraints, (2.17b) and (2.17c), the one that requires a higher transmission power is satisfied with equality, we can rewrite the problem as follows.

$$\min_{\mathbf{P}_0^m \succeq 0} \quad 0 \tag{2.18a}$$

$$\text{s.t.} \quad \eta_{u,n}^m = q_{u,n}^m, \forall n \in \mathcal{N}^M, \forall u \in \mathcal{U}_0^m, \tag{2.18b}$$

where $q_{u,n}^m = \max\{\eta_{thr} \cdot X_{u,n}^m, 2^{\tilde{r}_u^m \cdot X_{u,n}^m} - 1\}$. The problem is always feasible by assumption, and the optimal power for each active channel is given by solving the equality constraint Eq. (2.18b). Therefore, we can find the power update method directly from P. 2.18 after plugging in the SINR expression, Eq. (2.5a), to $\eta_{u,n}^m$. Then, the distributed power update method for an active channel n of MBS m which is associated with UE u becomes: $P_{0,n}^m[t+1] = \frac{q_{u,n}^m}{G_{u,n}^{0,m}}(\sum_{m' \neq m \in \mathcal{M}} G_{u,n}^{0,m'} \cdot P_{0,n}^{m'}[t] + \Delta f \cdot N_0)$. It is worth mentioning that this is equivalent to the power update method in Eq. (2.13) provided the feasible power region is nonempty and both the channel gain and the amount of interference are reported from UE.

According to [48], an iterative power update method, in general, is given by $\mathbf{p}[t+1] = \mathbf{I}(\mathbf{p}[t])$, where $\mathbf{I}(\cdot)$ is *interference function*.[6] We use $I_{0,n}^m(\cdot)$ to indicate the interference function for an active channel n of MBS m. The interference function is *standard* if it satisfies *positivity*, *monotonicity* and *scalability* properties for all nonnegative power vectors. Also, we use an overloaded notation $\mathbf{P}_0$ to indicate the transmission power of all MBSs. The positivity property, $I_{0,n}^m(\mathbf{P}_0) > 0$, is always satisfied because of the strictly positive background noise—even when $\mathbf{P}_0 = \mathbf{0}$, we have $\frac{q_{u,n}^m}{G_{u,n}^{0,n}}\Delta f \cdot N_0 > 0$. The interference function also satisfies the monotonicity property, i.e., if $\mathbf{P}_0^+ \succeq \mathbf{P}_0$, then $I_{0,n}^m(\mathbf{P}_0^+) \geq I_{0,n}^m(\mathbf{P}_0)$. Let $\mathbf{P}_0^+ = (1+\epsilon)\mathbf{P}_0$ for $\epsilon \geq 0$. Then, we have

$$\begin{aligned}
I_{0,n}^m(\mathbf{P}_0^+) &= I_{0,n}^m((1+\epsilon)\mathbf{P}_0) \\
&= I_{0,n}^m(\mathbf{P}_0) + \frac{q_{u,n}^m}{G_{u,n}^{0,m}}(\epsilon \sum_{m' \neq m \in \mathcal{M}} G_{u,n}^{0,m'} \cdot P_{0,n}^{m'}) \\
&\geq I_{0,n}^m(\mathbf{P}_0),
\end{aligned}$$

[6] Note that the notation $\mathbf{I}$ in Section 2.3.5 is different from the one in Section 2.3.3.

from which we can conclude that the monotonicity property is always satisfied. Finally, the positivity property and convexity of the interference function imply scalability, i.e., for all $\alpha > 1$, $\alpha I_{0,n}^m(\mathbf{P}_0) > I_{0,n}^m(\alpha \mathbf{P}_0)$.

Since the interference function $I_{0,n}^m(\cdot)$ satisfies the three properties, the proposed power update method is called *standard power control algorithm* [48]. Due to the convexity of the problem P. 2.17 (or P. 2.6), there exists an optimal power allocation vector, meaning that the proposed power update method has a fixed point. Then, the fixed point is unique by [48, Theorem 1]. Finally, by using [48, Theorem 2] we conclude that the proposed power update method converges to a unique fixed point for any initial power vector as long as the feasible power region is not empty. ■

2.3.6 Overall Procedure

In this section, the overall procedures of the proposed scheme is given, i.e., the BS association and channel assignment in Stage 1 and the power allocation in Stage 2, as a summary of the current section. At the beginning of Stage 1, all BSs transmit pilot signals over the entire channels to which they have an access. UE u senses the signal, calculates the per-channel cost, and transmits the cost vector $\mathbf{c}_u^m$ to MBS m. Then, MBS m determines the UE-BS association and channel assignment by running Algo. 1. The decision made by MBS m is broadcasted to all SBSs ($\forall s \in \mathcal{S}^m$) and all UEs ($\forall u \in \mathcal{U}^m$). Each active MBS channel manager with an associated UE runs Eq. (2.13) to determine the downlink power, and the UE sends the measured channel gain and interference back to the channel manager. This power allocation procedure iterates until the change of the power becomes less than the given threshold.Each active SBS channel manager with an associate UE runs the same procedure except that it runs Eq. (2.16) to determine the downlink transmission power. While each active channel manager tries to determine the transmission power, each BS checks if the sum power exceeds the power budget. If it does, the BS stops all active channel managers and lets them use the same power, $P_{0,n}^m = P_{max}^M / |\mathcal{U}_0^m|$ (in case of MBS) and $P_{s,n}^m = P_{max}^S / |\mathcal{U}_s^m|$ (in case of SBS) for downlink communication. If it does not, the BS waits until all active channel managers finish their power allocation procedures.

2.4 Evaluation

We implemented and simulated the proposed algorithm along with others for comparison on top of MATLAB [49] and CVX [50]. The following Section 2.4.1 describes the network configurations and parameter settings which are common to all scenarios considered in this section. In Section 2.4.2 we show that for Stage 1, the optimality gap between the proposed solution (i.e., Algo. 1) and the optimal BS association and channel assignment (i.e., P. 2.2) is small. What follows is the performance evaluation and comparison of the proposed scheme to others in terms of the power consumption on different networks, i.e., single-cell, small-scale and large-scale networks, in Section 2.4.3, Section 2.4.4 and Section 2.4.5, respectively.

For comparison to the optimal solution, we introduce a new metric, $D_X(\cdot)$, to measure the difference in the Stage 1 decision between a certain method and the optimal solution, which is defined as follows: $D_X(\texttt{<method>}) = ||X^*_{optimal} - X^*_{method}||_1$, where $||Y||_1 = \sum_{\forall y \in Y} |y|$. Here, $X^*_{optimal}$ is the optimal solution found by solving P. 2.2, whereas X^*_{method} is the optimal solution found by solving the corresponding problem for `<method>`. In other word, $D_X(\texttt{<method>})$ counts the number of entries that do not match between the two solutions.

In addition, we have implemented one more scheme, called SSSF (Strongest Signal Strength First) for comparison. In contrast to the proposed method which considers both the channel gain and the service demand, SSSF takes only the signal strength into account when making BS association and channel assignment. It is easy to implement SSSF or any similar variations due to the general structure of the Stage 1 problem. In contrast to the proposed method which minimizes the cost values, SSSF maximizes the sum of the benefit which is equal to the channel gains. After replacing $\mathbf{c}^m$ with the benefit, we can simply replace the objective function P. 2.2 with the benefit-sum maximization problem. We have directly solved SSSF by using the MATLAB (M)ILP solver which uses Branch-and-Bound algorithm.

2.4.1 Network Configuration

There are M MBSs that are regularly deployed with keeping the inter-cell distance of 600 m among adjacent ones. Each MBS has 300 m of coverage, and is overlaid by S indoor SBSs and U UEs. We have used a Uniform distribution for locating SBSs and UEs. A UE is located indoor if it is placed within the coverage of a SBS which is 30 m. The QoS requirement of each UE is randomly drawn from a Uniform distribution. The total BS transmission power of MBS and SBS is 46 dBm (40 W) and 20 dBm (100 mW), respectively [42].

The channel model from [42] is used, which includes the distance dependent path-loss, penetration loss (when applicable), multipath fading and lognormal shadowing. The path-loss between a BS and a UE is listed below. The unit of path-loss is dB, and R is the distance between two entities in the unit of meter.

- MBS and an indoor UE: $15.3 + 37.6\log(R) + L_{ow}$,
- MBS and an outdoor UE: $15.3 + 37.6\log(R)$,
- SBS and its associated UE: $38.46 + 20\log(R)$, and
- SBS and an outdoor UE:
  $\max\{38.46 + 20\log(R), 15.3 + 37.6\log(R)\} + L_{ow}$,

where L_{ow} is the penetration loss of an outdoor wall, which is 20 dB. In case of the path-loss between an indoor UE and an SBS located in a different building, the penetration loss gets doubled. The Rayleigh fading model is used to capture the multipath effect, and the standard deviation of lognormal shadowing is as follows.

- MBS and an indoor UE: 10 dB,
- MBS and an outdoor UE: 10 dB,
- SBS and its associated UE: 4 dB, and
- SBS and an outdoor UE: 8 dB.

In addition, for a fair comparison to [36] we set $\Delta f = 180$ kHz and per-Hz noise power $N_0 = 10^{-13}$ W in accordance with the parameters declared therein.

2.4.2 Optimality Gap in Stage 1

As aforementioned in Section 2.3.3, the Algo. 1 iteratively solves P. 2.3 and recovers binary solutions instead of directly solving P. 2.2 to lower the computational complexity. Due to the relaxation on binary variables, Algo. 1 may yield a suboptimal solution to P. 2.2 which possibly affects the power consumption in the subsequent Stage 2 for power allocation. In order to check by how much the solution of Algo. 1 is deviated from the optimal solution to P. 2.2, we have implemented and solved P. 2.2 by using the MATLAB (M)ILP solver which uses Branch-and-Bound algorithm. Each data point in both Fig. 2.1 and Fig. 2.2 is an average of 20 runs of randomly-generated scenarios, where there are 4 SBSs on an MBS. The number of UEs in a macrocell is set to 10, 20, $\cdots$, 50. Also, 95% of the confidence interval is marked on each data point in Fig. 2.1.

Fig. 2.1 shows the (normalized) minimum cost found by running Algo. 1 and P. 2.2, referred to as Proposed and Optimal, respectively, in the figure. For each different number of UEs on a macrocell, the proposed method results in a close-to-optimal objective value. In order to take a closer look at the difference in the two objective values, Fig. 2.2 shows the error ratio $e = |\hat{p}^* - p^*|/p^*$, where $\hat{p}^*$ and p^* are the normalized minimum cost found by running Algo. 1 and solving P. 2.2, respectively. As it can be seen in Fig. 2.2 the error ratio becomes stable as the number of UEs increases and does not exceed 0.008. That is, the proposed Algo. 1 yields a sub-optimal solution to P. 2.2 with a small optimality gap.

In what follows, we show the power consumption of the proposed method along with others for comparison, and show the effect of the sub-optimality on the power consumption.

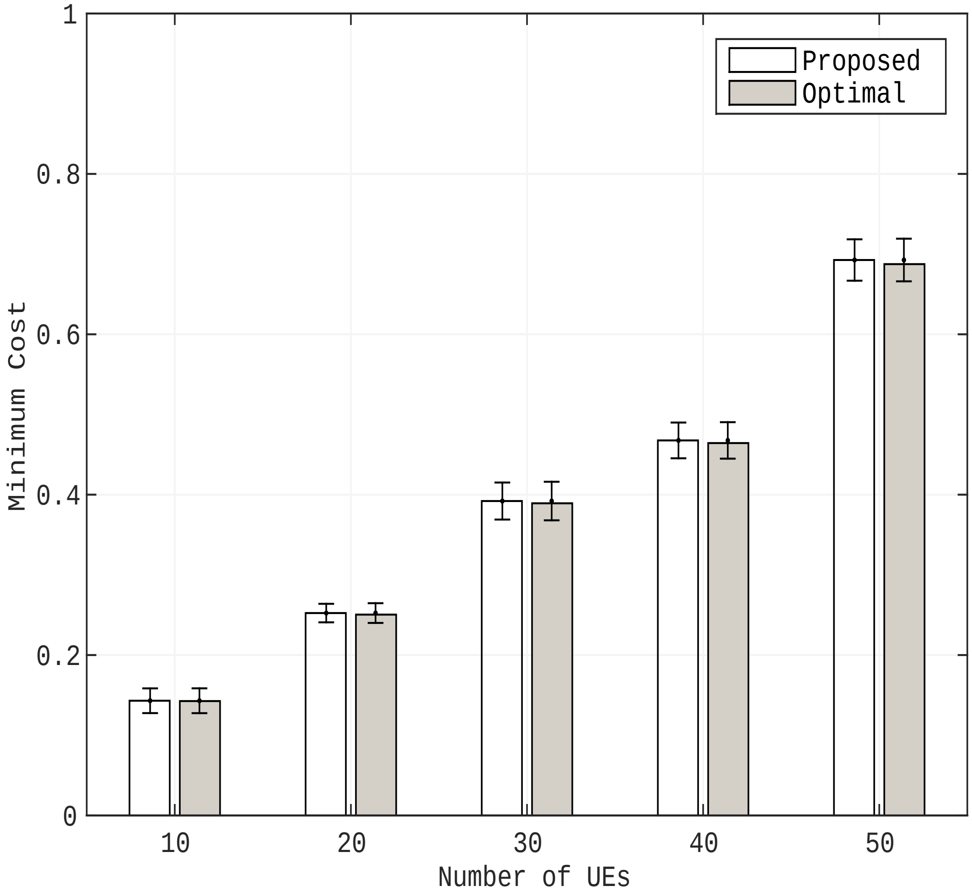

Figure 2.1: Minimum cost in Stage 1 for the optimal and the proposed BS/channel assignment method.

2.4.3 Single-Cell Networks

In addition to comparing to the optimal solution and SSSF, we compare the performance of the proposed method to [36], which is denoted by Abdelnasser in Fig. 2.4 and Fig. 2.5. In contrast to the proposed scheme which assumes an independent channel deployment between different types of BSs, [36] shares all available channels between an MBS and all SBSs therein, called co-channel deployment. The proposed resource scheduling scheme in this chapter allocates an optimal power to each channel by taking both the channel gain and the QoS requirement of an associated UE into account, while [36] allocates an equal power to all channels in use. The proposed association

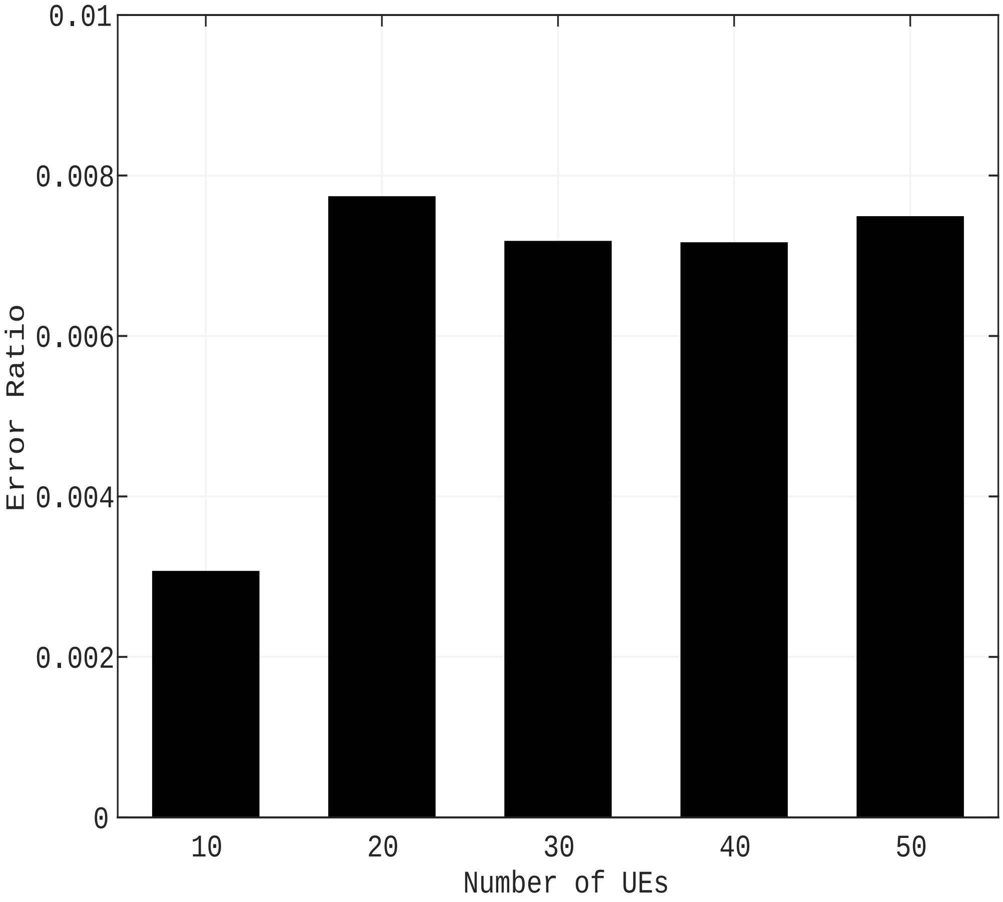

Figure 2.2: Error ratio of the minimum cost in Stage 1 for the proposed BS/channel assignment method.

scheme in this chapter allows UEs to be dynamically offloaded to SBSs for capacity enhancement (i.e., open access mode), whereas [36] does not (i.e., closed access mode). Since the work in [36] considers only a single MBS, we set up a similar network as Fig. 2.3, where there is a single MBS on the network with 4 SBSs and 20 UEs therein. Please note that due to this limitation, [36] is not used for performance comparison in the following Section 2.4.4 and Section 2.4.5. In this simulation, we have $D_X(\texttt{proposed}) = 0$ and $D_X(\texttt{SSSF}) = 6$.

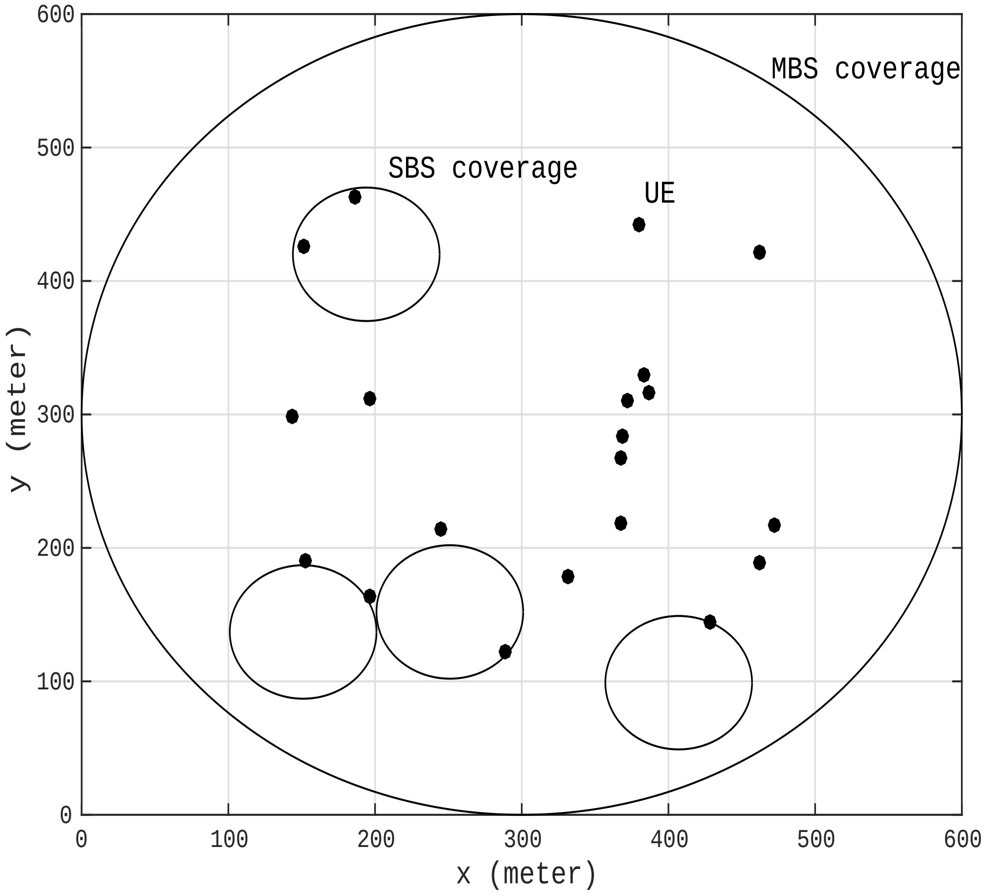

Figure 2.3: Network scenario for a single-cell network.

2.4.3.1 Power Consumption

Fig. 2.4 shows the overall power consumption, i.e., the total power used by an MBS and SBSs on the network, for the four different schemes. As it can be seen in Fig. 2.4, the work in [36] uses more power than the proposed method as well as Optimal and SSSF. In contrast to the proposed method in this chapter that dynamically manages the interference, [36] takes a *conservative* approach. In [36], when a UE is associated with an MBS, the MBS calculates the maximum allowable interference on the allocated channel for the UE, and then assigns the maximum power to the channel which is P_{max}^{M} divided by the number of channels in use. On the other hand, the proposed scheme in

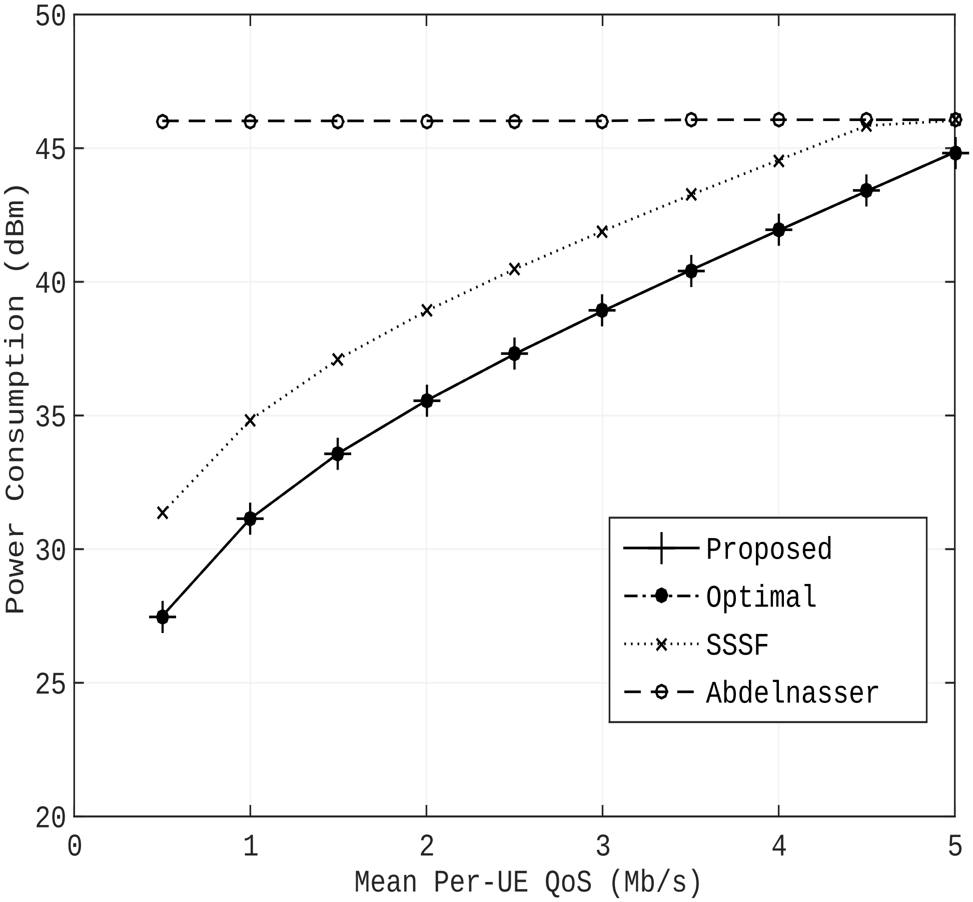

Figure 2.4: Overall power consumption for a single-cell network.

this chapter as well as both Optimal and SSSF allocates the minimum power to each channel while satisfying both SINR and QoS requirements for the associated UE. Therefore, our proposed work much outperforms [36] in terms of the power consumption especially when the amount of downlink traffic is small. Since we have $D_X(\texttt{proposed}) = 0$, the proposed scheme results in the optimal solution. SSSF is also able to dynamically adjust the transmission power, and thus, its power usage gradually increases as the average service demand increases. However, our proposed scheme consumes less power than SSSF. That is, considering both channel gain and QoS requirements results in a more power-efficient solution than taking only the signal strength into account.

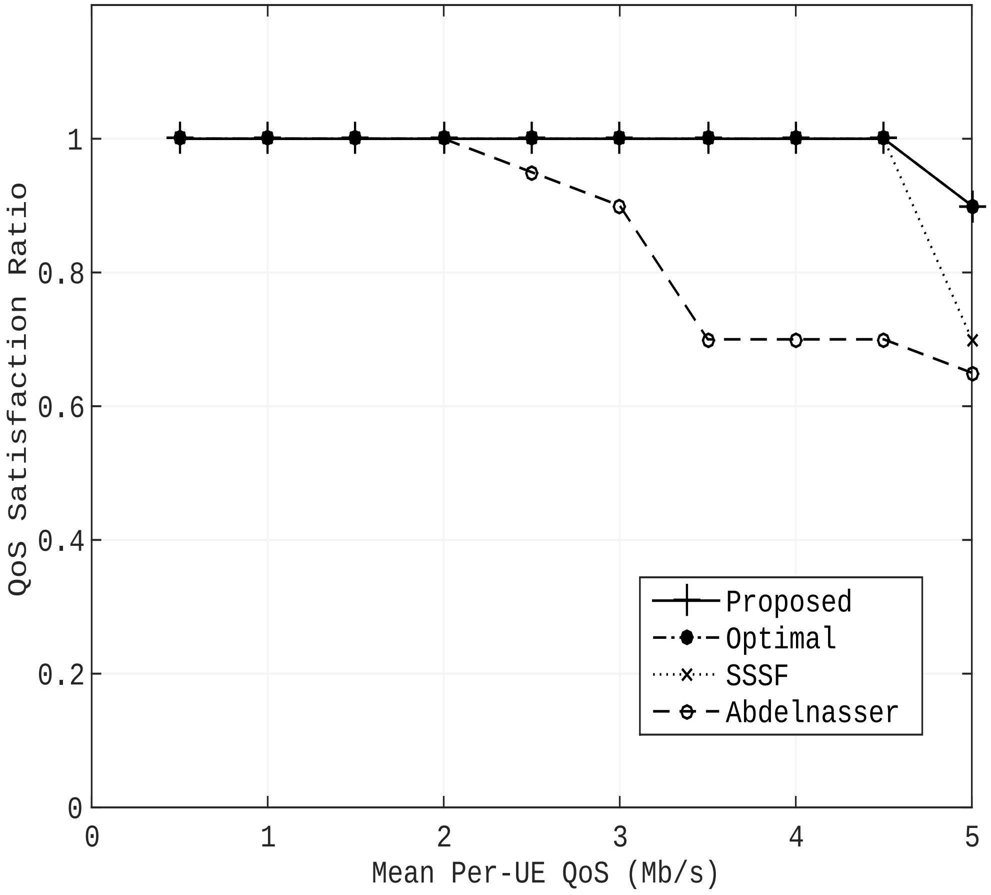

Figure 2.5: QoS satisfaction ratio for a single-cell network.

2.4.3.2 QoS Satisfaction Ratio

The Fig. 2.5 shows the QoS satisfaction ratio which is the ratio of the number of UEs with their QoS satisfied to the total number of active UEs on the network. As it can be seen in the figure, the satisfaction ratio of the work in [36] starts to drop when the mean QoS becomes larger than 2 Mb/s. On the other hand, the other methods successfully satisfy the QoS requirements of all UEs until the mean QoS is 4.5 Mb/s. Due to the lack of freedom in controlling the downlink transmission power, the work in [36] always allocates the fixed transmission power to all active channels. This inflexibility in power allocation may not be efficient, because it allocates more than necessary amount of power to the UEs with high channel gains, while failing to satisfy the QoS requirement of UEs with low gains. The proposed method and the Optimal, on the other hand, has a higher level of freedom in power control than [36], because each channel manager allocates the minimum power level that satisfies both SINR and QoS requirements of the associated UE. When the average per-UE QoS is 5 Mb/s, the QoS satisfaction ratio of SSSF drops sharply, while it is not the case for both the proposed and optimal scheme even though both experience a small amount of degradation.

2.4.4 Small-Scale Networks

We have evaluated the proposed method on a small-sized network where there are 3 MBSs on the network as shown in Fig. 2.6. The locations of these three MBSs form an equilateral triangle, meaning that the distance from any MBS to either of the rest two is the same. Each MBS is overlaid by 4 SBSs and 20 UEs. In this simulation, we have $D_X(\texttt{proposed}) = 0$ and $D_X(\texttt{SSSF}) = 18$. Therefore, the performance of the proposed scheme will be exactly same as that of Optimal.

2.4.4.1 Power Consumption

The Fig. 2.7 shows the power consumption of the three methods, the proposed, optimal and SSSF, with respect to different mean per-UE QoS. The overall power use is the sum of transmission power used by all macro and smallcell BSs on the network. Although it is not shown in the figure, the difference between the overall power use and the aggregate MBS power use is trivial, meaning that MBSs use most of the power consumed in the network. This is because an MBS associates with much larger number of UEs than SBSs due to the long transmission range and a larger power budget. Thus, a UE associated with an MBS may have a small channel gain, making the MBS use a high transmission power to satisfy the UE's SINR and QoS requirement. On the other hand, an SBS has a small number of associated UEs with a short distance to each. Therefore, it does not need much power to satisfy the associated UE's QoS demand and SINR requirement. The overall

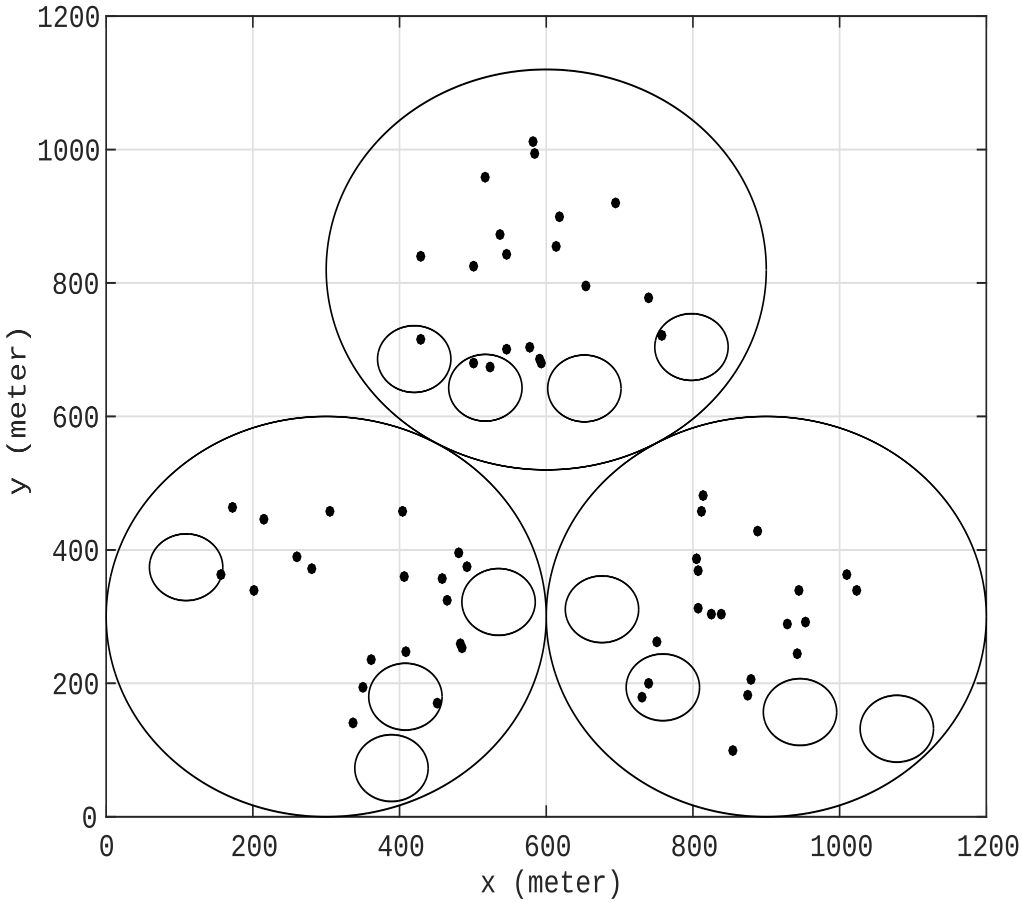

Figure 2.6: Network scenario for a small-scale network.

power consumption becomes saturated when the mean per-UE QoS demand is approximately 5 and 6 Mb/s, respectively, for SSSF and both the proposed and optimal schemes.

2.4.4.2 QoS Satisfaction Ratio

QoS satisfaction ratio is the number of UEs with their QoS satisfied to the total number of active UEs on the network. As it can be seen in Fig. 2.8, the QoS requirements of all UEs are fully satisfied when the mean per-UE QoS is equal to or less than 4 or 4.5 Mb/s, respectively, for SSSF or both the proposed and optimal. Then, the QoS satisfaction ratio drops as the per-UE demand becomes larger. It is noteworthy that for the first one or two drops of the ratio, SSSF

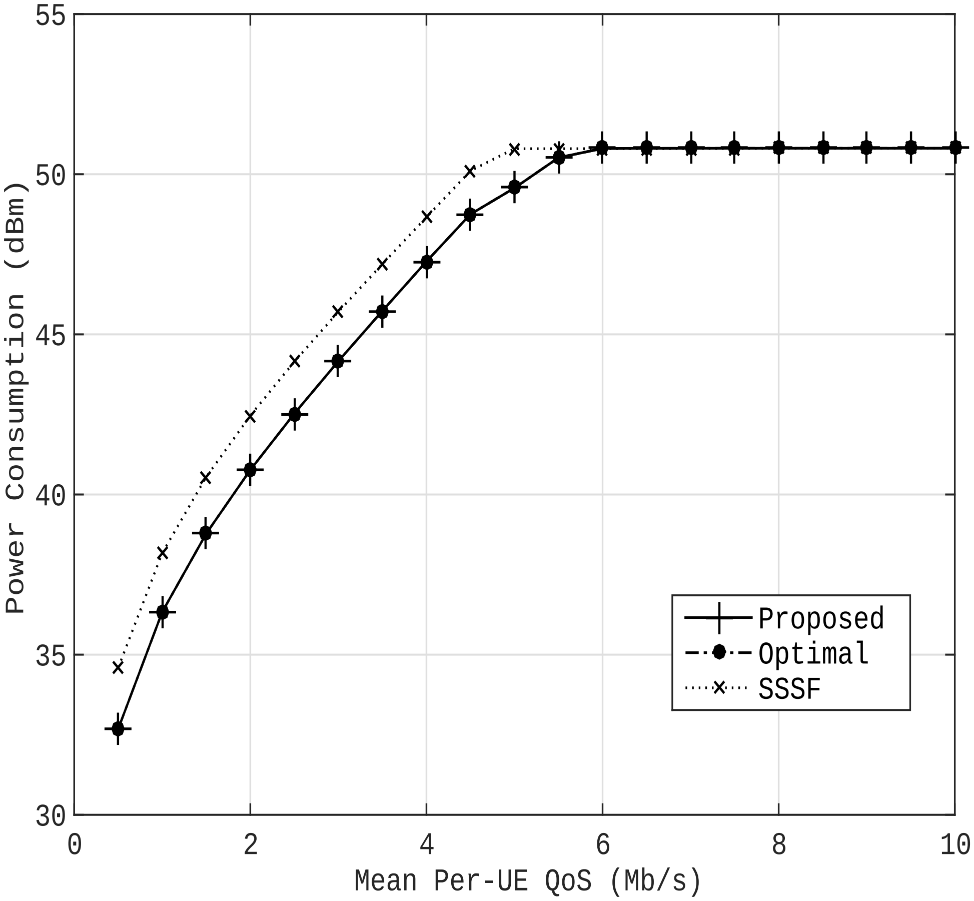

Figure 2.7: Overall power consumption for a small-scale network.

shows a steeper decline than the rest two. Since it already uses much power when the mean per-UE QoS is 4 Mb/s, further increase in QoS causes a significant drops in the QoS satisfaction ratio. The increase in the QoS requirement will eventually let all BSs use the maximum transmission power, which increases the interference level. Failing in achieving the satisfaction ratio of 1 means there is at least one BS whose total power budget constraint is violated. Note that the violation of the power budget constraint makes a BS use an equal power allocation for all active channels. Since the equal power assignment takes away the freedom in power control from a BS, any further increase of QoS requirements will yield more UEs with their QoS unsatisfied. UEs in a SSSF network suffer

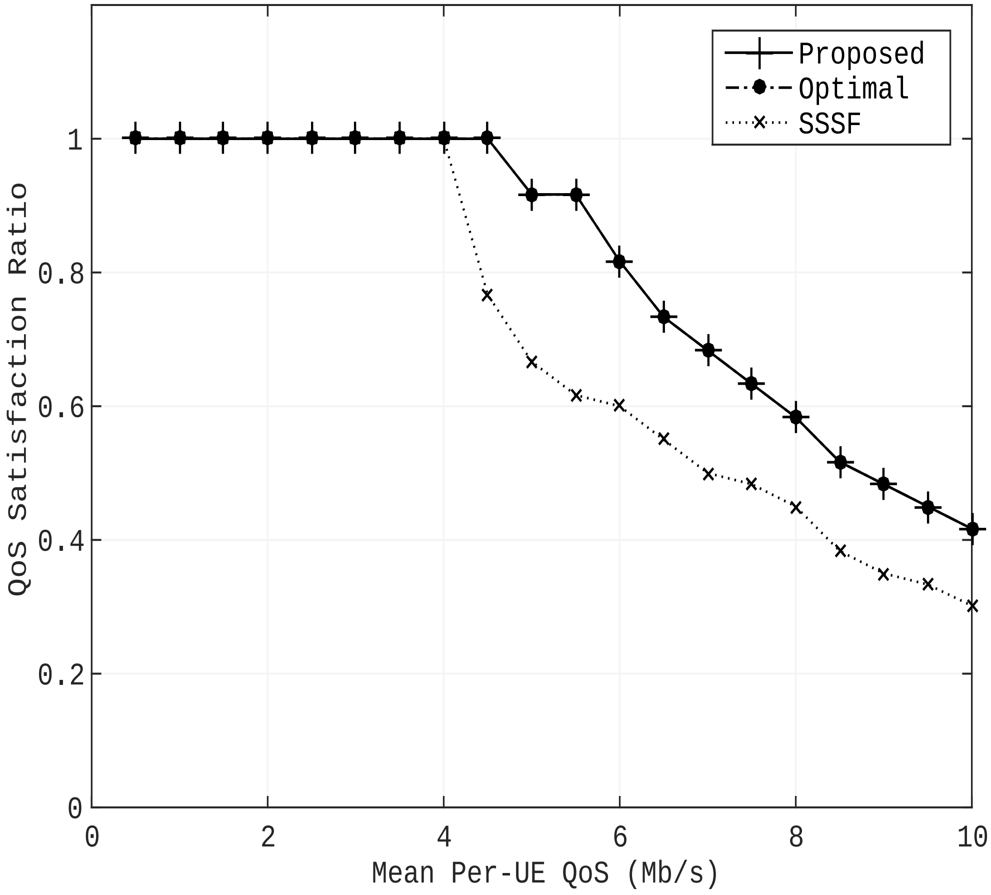

Figure 2.8: QoS satisfaction ratio for a small-scale network.

much more QoS degradation as the average QoS demand increases compared to both the proposed and the optimal schemes.

2.4.4.3 Convergence

The speed of convergence determines whether the proposed algorithm is suitable for an online processing or not. We have evaluated the speed of convergence with two different setting as follows, while keeping the rest network configurations the same. One is *synchronized* power update and the other is *unsynchronized*. In the synchronized power update setting, all active channel managers update their power at the same time. In other words, on iteration t it is guaranteed that all

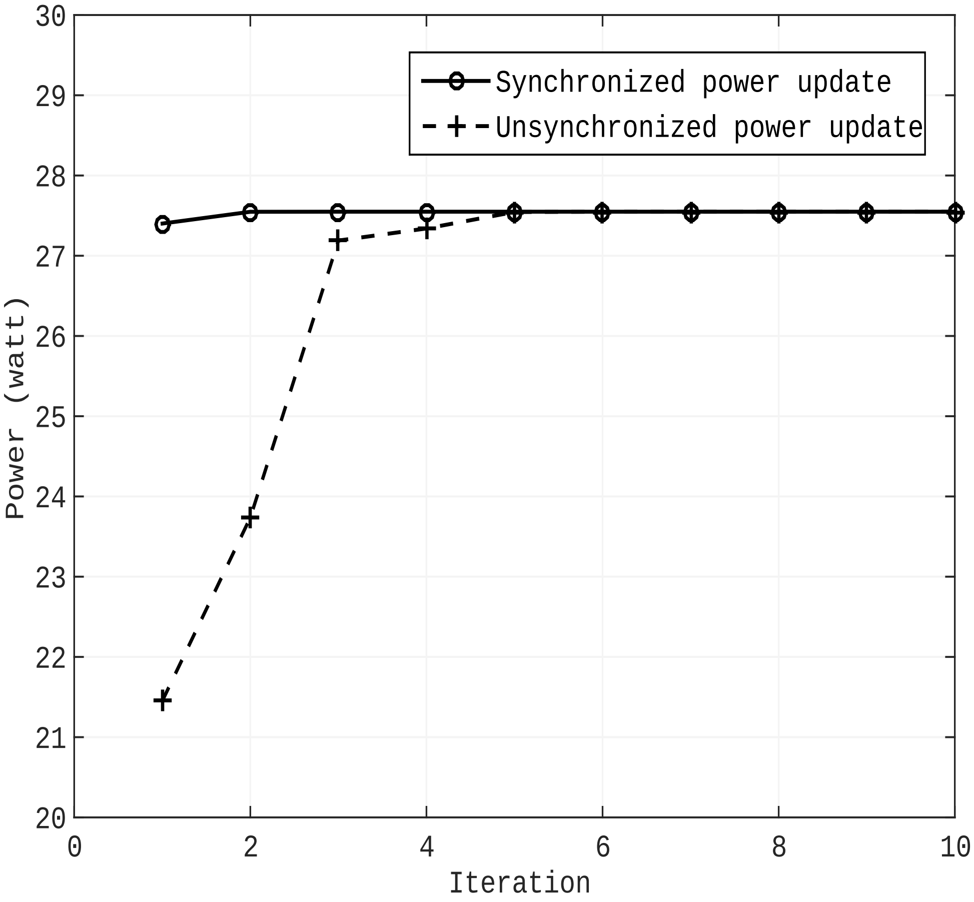

Figure 2.9: Convergence of the iterative power allocation procedure for a small-scale network.

the active channel managers on the network have fished their power update procedures for the previous $t-1^{st}$ iteration. On the other hand, the unsynchronized power update does not assume the synchronized power update. For example, when a channel manager runs its t^{th} power update, it is possible that there are some channel managers that have not finished their $t-1$ iterations (or even the ones before the $t-1^{st}$ iteration). In order to emulate the unsynchronized updates, we have added some random delay into the power update method. The Fig. 2.9 shows the convergence behaviors of the power update method under different settings. As expected, the unsynchronized setting requires more iterations. In this small-scale network, every MBS has two effective interfering cells,[7] both of which are 600 m away. Therefore, when the update procedures are synchronized, all BSs will converge at the same time on a small-scale network. Note that this is not the case for a large-scale network where each cell is exposed to the different number of effective interfering cells.

2.4.5 Large-Scale Networks

We also carried out an evaluation on a large-scale network where there are 30 MBSs, each of which is overlaid by 4 SBSs and 20 UEs. MBSs are placed in a shape of a bee hive where there are 5 rows of MBSs with 6 MBSs per row as in Fig. 2.10. The rest configurations remain the same as before. In this simulation, we have $D_X(\texttt{proposed}) = 22$ and $D_X(\texttt{SSSF}) = 358$. The proposed method has failed to find the optimal solution in BS association and channel assignment for this network for having a nonzero value of $D_X(\texttt{proposed})$. However, since the difference is not significant, the power consumption in Stage 2 is expected not to be deviated much from that of the optimal method.

2.4.5.1 Power Consumption

The power consumption of BSs on a large-scale network is illustrated in Fig. 2.11, which has a similar trend to Fig. 2.7. It is worth mentioning that although the proposed method has a (trivially) different Stage 1 solution than that of the optimal method, it is hardly seen on the power consumption. In both small- and large-scale cases, the overall power use of both the proposed and optimal becomes saturated around 6 Mb/s of the mean QoS. Also, both methods outperform SSSF in terms of power consumption, indicting that considering both channel gain and QoS requirement is more energy-efficient than the one considering only the received signal strength.

2.4.5.2 QoS Satisfaction Ratio

Each cell on the large-scale network experiences more interference than that on the small-sized network for the increased number of effective interfering cells. Fig. 2.12 shows the QoS satisfaction

[7] A cell is an effective interfering cell to another if the interfering cell is close enough to the interfered cell. For example, if an interfering cell is 600 m away, it is an effective interfering cell. However, a cell which is 600 km away is not an effective interfering cell, because the interference from the cell is too weak.

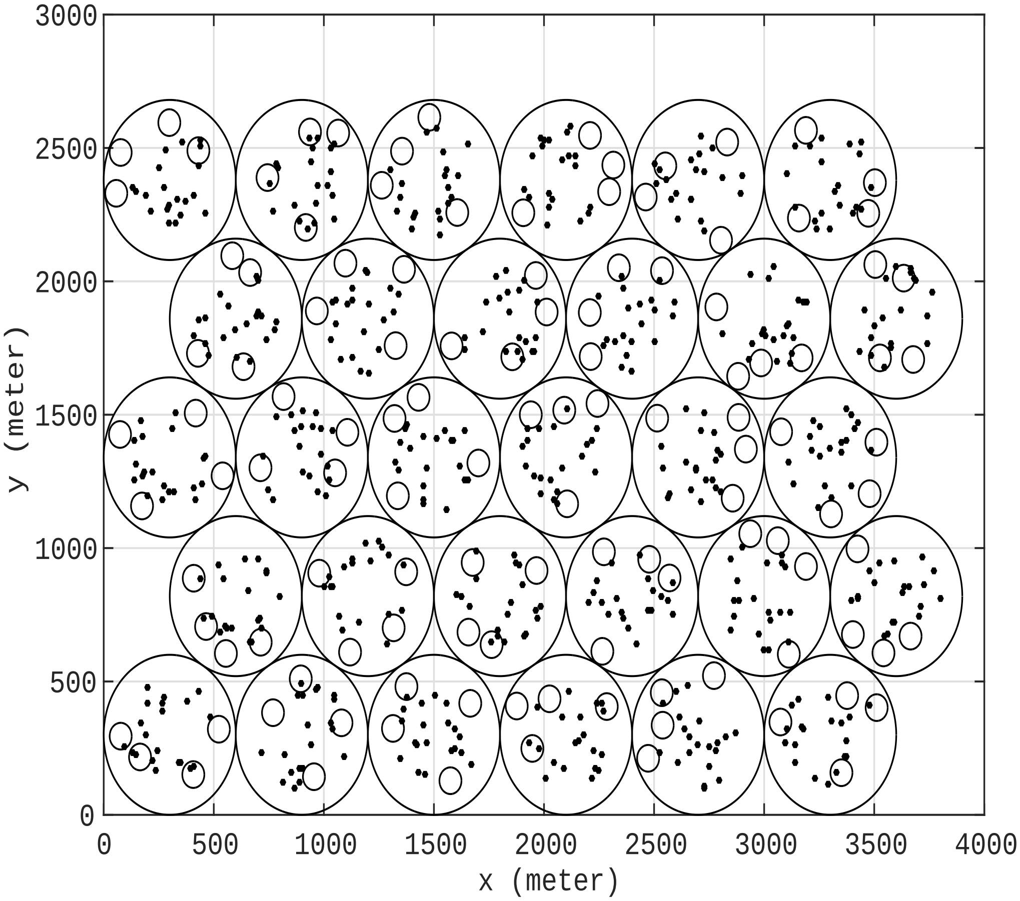

Figure 2.10: Network scenario for a large-scale network.

ratio of UEs on the large-sized network. As it can be seen in the figure, the ratio starts to drop when the mean QoS becomes larger than 2 and 3.5 Mb/s, respectively, for SSSF and both the proposed and optimal, which was not the case on the small-scale network. This is mainly because of the increased level of interference on the large-scale network. Having more effective interfering cells on the network increases the level of interference to each cell, and thus results in a lower energy efficiency. The performance degradation of the proposed method for having a non-optimal solution in Stage 1 can be seen in Fig. 2.12. As the average QoS requirement becomes larger than 8 Mb/s,

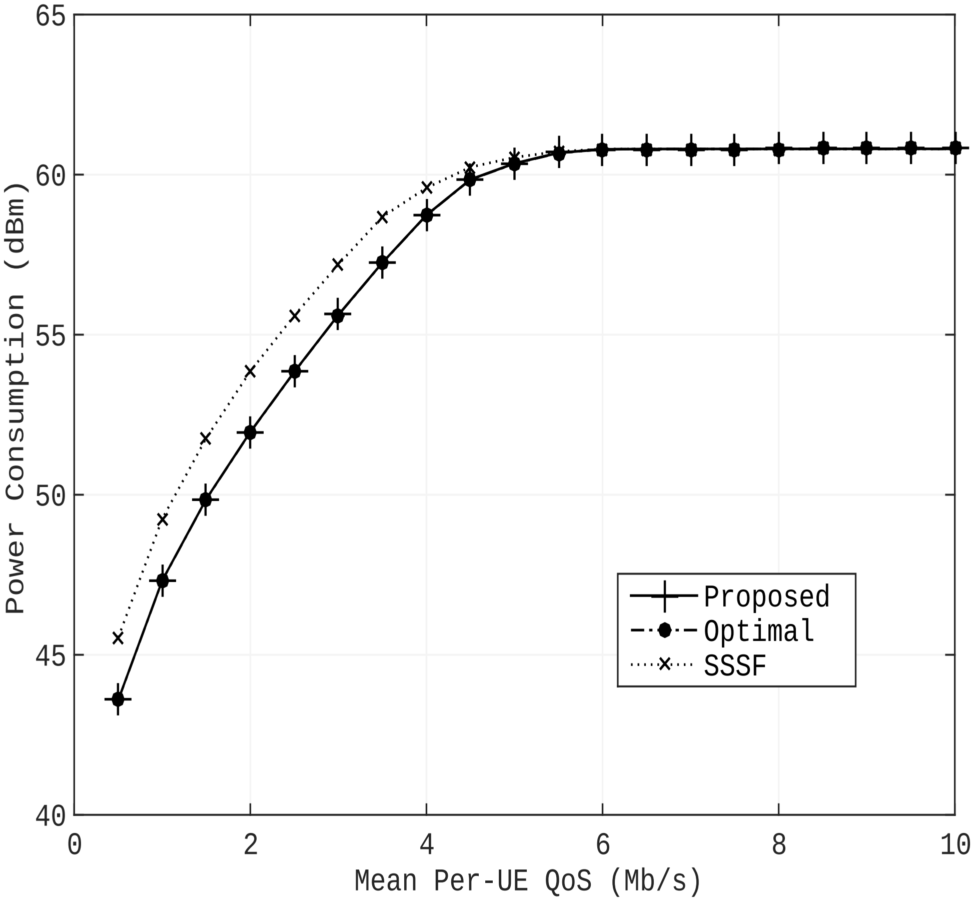

Figure 2.11: Overall power consumption for a large-scale network.

the proposed method results in a slightly lower QoS satisfaction ratio than the optimal method, but the degradation is trivial.

2.4.5.3 Convergence

Due to the increased number of interfering cells, the iterative power update procedure on the large-scale network takes a couple of more steps to converge as shown in Fig. 2.13. Still, the network reaches convergence in 6 or 7 iterations even when the power update is not synchronized. This fast convergence behavior on the large-scale network proves that the proposed scheme scales well with the network size, making it suitable for an online resource scheduling.

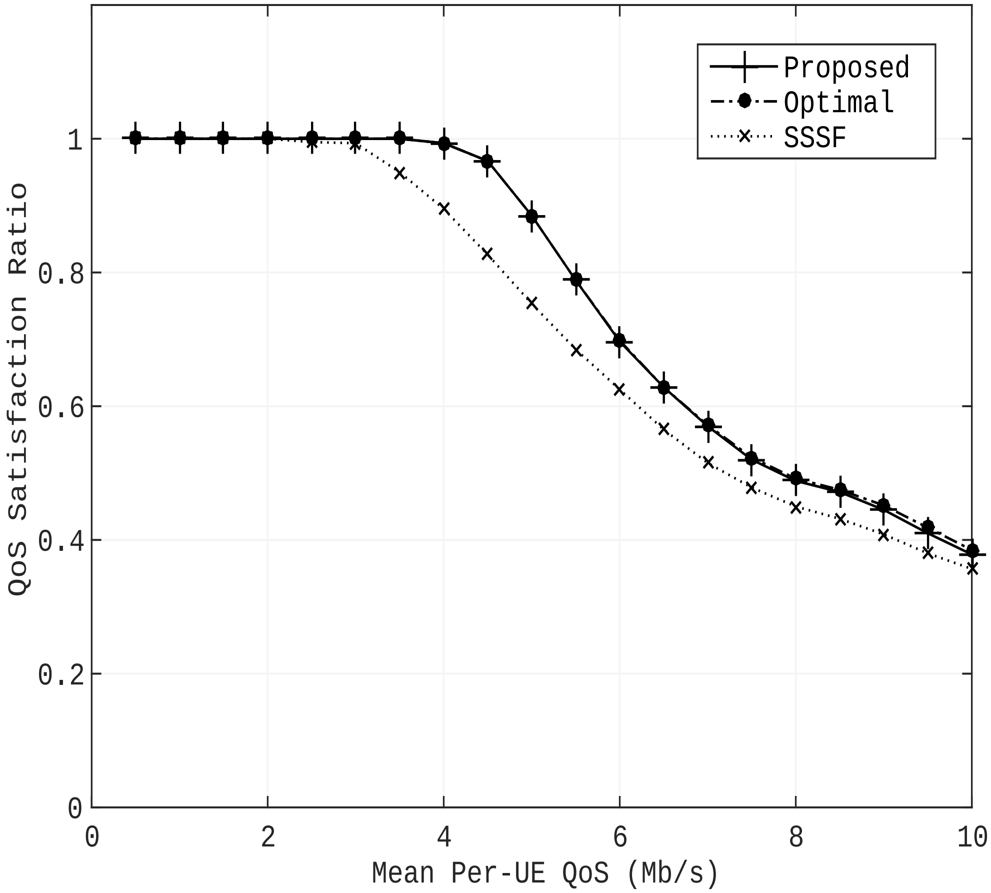

Figure 2.12: QoS satisfaction ratio for a large-scale network.

2.5 Conclusion

In this chapter, we have proposed a distributed and energy-efficient association and resource scheduling scheme for two-tier HetNets. We have formulated an optimal user association problem and then proposed a resource allocation algorithm in such a way that they can be solved efficiently by an iterative, distributed method. To be specific, we have formulated an optimal user association and channel assignment problem for Stage 1 and then applied a relaxation and an iterative adjustment method so as to make the problem tractable and low-complex. In addition, we have transformed the proposed power assignment problem into a set of lightweight distributed proce-

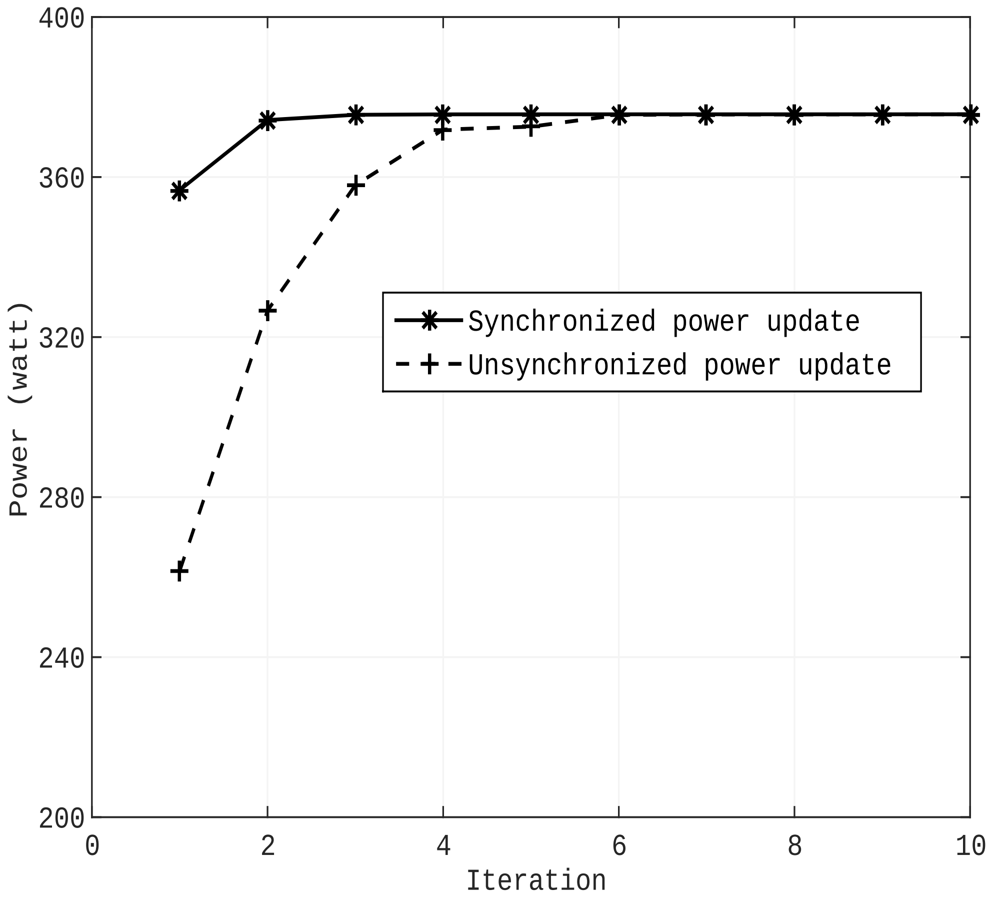

Figure 2.13: Convergence of the iterative power allocation procedure for a large-scale network.

dures by using the decomposition structure for Stage 2. The comparison results and the evaluation studies on small-/large-scale networks show that the proposed scheme maintains a low power consumption while satisfying users' QoS requirements with a low computational load, which proves that the proposed scheme can be used for an online resource scheduling for HetNets.

CHAPTER 3. COST AND ENERGY-EFFICIENT RESOURCE OPTIMIZATION FOR C-RAN BASED NEXT GENERATION CELLULAR NETWORKS

3.1 Summary

As network operators invest more and more in infrastructure to keep up with the ever-increasing traffic demand, it has become important for them to operate networks in such a way that they can maximize profit while minimizing cost. Such an expansion in network facilities also increases power consumption which has a negative impact on both environment and revenue. In this regard, the cloud radio access network (C-RAN) which is a promising next-generation network architecture has gained much attention. Recently, the centralized computing and being capable of resource re-configuration/virtualization help to schedule the network resource in a cost-effective manner. In this chapter, we study an optimal resource allocation for C-RANs to maximize profit while minimizing power consumption. Moreover, by allowing network operators to share their resources through traffic offloading, we show that the profit can be further increased. The proposed multi-stage stochastic programming model makes a robust optimal decision that effectively responds to uncertainties from users' mobility and service demand. Evaluation and comparison results show that the proposed approach is highly cost-effective under network uncertainties and with traffic offloading.

3.2 Introduction

As we move towards the next generation network (5G), we have seen some meaningful progress made for it. Among those, the centralized/cloud radio access network (C-RAN) [51] [52] which is a recently-proposed cellular network architecture is considered to be one of the key enablers for 5G. As shown in Fig. 3.1, C-RAN consists of a BBU (baseband unit) pool, RRHs (remote radio heads)

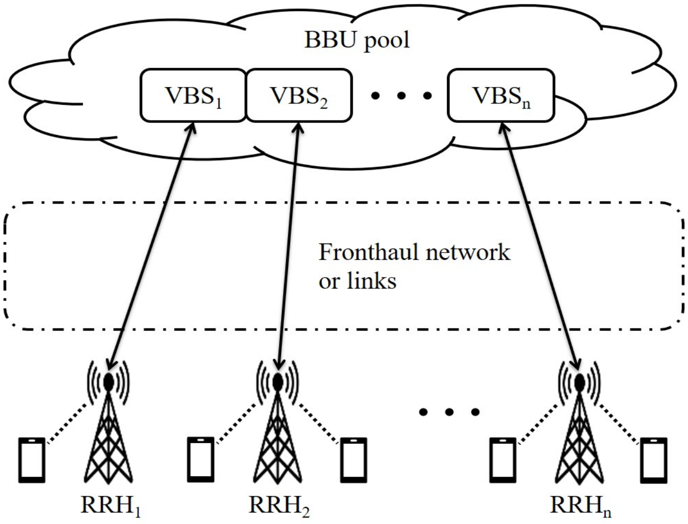

Figure 3.1: C-RAN consists of BBU pool, RRHs and fronthaul links.

and fronthaul links that connect them via high-speed links such as optical transmission networks [53]. C-RAN shifts the majority of the functionalities from base stations (BSs) to the central cloud-computing resource pool (i.e., BBU), which is in contrast to the conventional belief that BSs are supposed to be in charge of MAC/PHY functions. To this end, the BBU pool in the cloud platform performs most of the tasks, e.g., baseband signal processing and transmission scheduling, making RRHs low-complex and low-cost. By virtue of the Software-Defined Radio and Network Function Virtualization, C-RANs can become more scalable, reusable and easily configurable. The centralization enables an enhanced cooperation/coordination between RRHs, and also the reduced cost of RRHs allows mobile network operators (MNOs) to densely deploy them (i.e., close to end-users) so that high transmission rates can be achieved with low transmission power. In the long run, C-RAN is expected to reduce the total cost of ownership, i.e., the sum of CAPEX (capital expenditure) and OPEX (operational expenditure), which makes it more attractive.

In particular, the heterogeneous C-RAN (H-CRAN) [54] which is an advancement to or variation on C-RAN is more attractive for the following reasons. By combining the C-RAN architecture with the conventional MBSs (macro base stations), H-CRANs can lead to a gradual transition from 4G to 5G. MBSs and RRHs, respectively, can be dedicated to handling control message exchanges and high-speed data transfer, the separation between the control plane and data plane becomes straightforward. In addition, due to the large coverage of MBSs, there is no need to densely deploy RRHs as shown in Fig. 3.5. In fact, highly-populated RRHs may cause control message overflow (e.g., due to frequent handovers [53]), increase the complexity of the interference management, and leave many of them under-utilized [55]. Such heterogeneous structure provides a new opportunity for resource optimization since it takes advantage of the centralized structure of C-RAN with low complexity due to the reduced number of RRHs.

In this chapter, we focus on how to operate H-CRAN in a profitable and energy-efficient manner by scheduling network resources under uncertainty. In contrast to the conventional approaches focusing on either throughput maximization or energy minimization, only a few studies have paid attention to profit maximization from an operator's perspective [56] [57] [58] [59]. It has recently been noticed that compared to the increasing investment that MNOs make to upgrade their network infrastructure, the average revenue per user has been small [53] [59]. Also, power consumption from the expanding infrastructure accounts for a large share of both increasing OPEX and carbon footprint from Information and Communication Technology. Thus, there is a urgent need for studying a cost-effective operation of the networking system in preparation for the upcoming 5G. In this regard, the overall goal in this chapter is to maximize the revenue, while minimizing the power consumption.

It is worth mentioning that the new structure of C-RAN and H-CRAN brings not only new opportunities but also new challenges. To effectively schedule the resources in C-RAN, the uncertainties brought by frequent changes in users location and their service demand must be carefully considered since both affect the overall performance to a large degree. To be specific, when there are short-range BSs deployed, the connectivity (or accessibility) between users and such BSs is largely affected by their geographic locations at the moment. In addition, varying users demand causes the traffic load in certain service area to change significantly [60] over time. Therefore, it is important to take both uncertainties into account at the same time for resource scheduling, which has not been carefully studied yet.

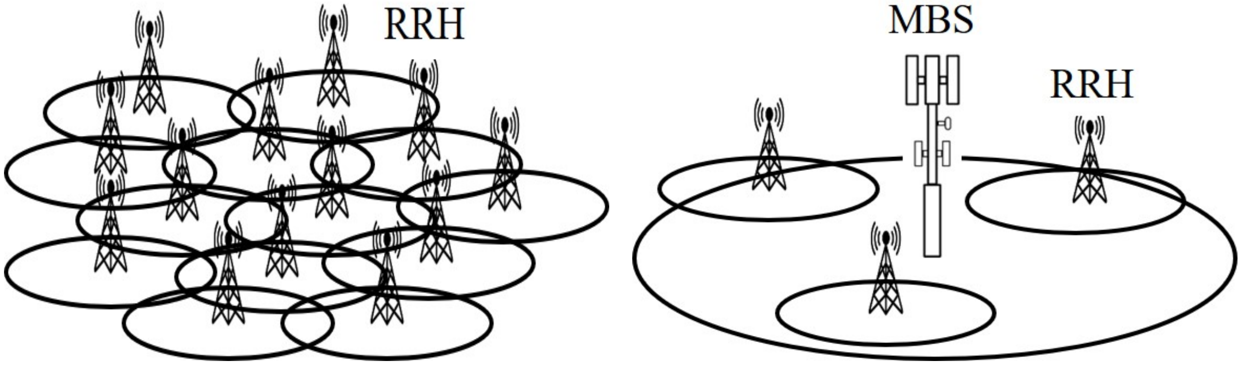

Figure 3.3 In C-RAN, RRHs are densely deployed to provide coverage, causing severe interference.

Figure 3.4 In H-CRAN, MBSs provide coverage, while heavy traffic is offloaded to RRHs.

Figure 3.5: Comparison between C-RAN and H-CRAN.

Another challenge in C-RAN is that some resources need to be treated separately from the rest in terms of the timescale of scheduling (or scheduling frequency). This is a critical issue since it might be different from what has been done thus far. For conventional cellular networking systems, it is important to optimally schedule networking resources such as channel assignment, transmission power, interference and user-BS association, and all of these are handled by the MBSs that UEs (user equipments)[1] are directly attached to at the moment. Since BSs makes decisions locally, online or dynamic resource allocation can be done without any significant delay. In C-RAN, on the other hand, Virtual Base Stations (VBSs) are built on the cloud platform by means of Virtual Machines (VMs), resulting in the increased system utilization, while decreasing CAPEX [61]. Considering that one or a set of VBSs on BBU is paired with an RRH to process the signals from/to it, frequent resizing of VBSs may bring a significant delay to traffic delivery[2]. In this regard, for the C-RAN resource scheduling, such resources need to be identified and separately handled from the rest that

[1] In this chapter, UE is used interchangeably with user.

[2] Re-configuring a VM, in general, is followed by adjusting a set of hardware resources, e.g., computing/processing power, disk storage, memory and network bandwidth, as well as installing an operating system and software packages so that a VM can be independently and fully functional. Also, an additional delay is introduced for the fronthaul link.

are applicable to online scheduling; otherwise, being unaware of such characteristic may yield an inefficient resource usage and service outage.

In this chapter, we propose an optimal resource scheduling method for H-CRAN under network uncertainties. In order for an MNO to maximize the profit, the proposed method optimizes a set of networking resources, including the partitioning of BBU pool for RRHs, the channel partitioning between different types of BSs, the association between BSs and users, and the channel assignment for each user. The fronthaul constraint and the power consumption are taken into account to make the proposed model to be more practical and energy-efficient, respectively. By using a stochastic programming (SP) approach, the proposed model can effectively react to the uncertainties from the users' mobility and their varying service demand. For the resources that are not suitable for frequent scheduling, the proposed model finds one-fits-all, robust solutions; for the rest resources, on the other hand, it finds an optimal solution in an online manner. Also, we propose a way to achieve a profitable cooperation between different MNOs by allowing users to perform traffic offloading (or roaming) between multiple MNOs [56] [57] [62] [63] when there is not enough resource left on their primary MNO's network. The evaluation studies compare the proposed model to others that do not or partially consider the uncertainties, and show that the proposed model can achieve a higher profit than the rest while minimizing both the power consumption and service outage ratio. Also we show the promising effect of the cooperation among different MNOs by means of traffic offloading on the profit.

In this study, we make the following major research contributions. First of all, we propose a cost and profit model with which we study how to maximize the overall profit. Such an approach has received little attention in the literatures. However, it is important in that compared to the conventional throughput maximization or energy minimization point of view, the profit maximization approach provides a perspective on how a network operator, i.e., the actual operator of the network, would schedule the resource use. Secondly, compared to the previous studies focusing on optimizing either the BBU pool and forming VBSs or the networking resources (e.g., associating UEs with MBS/RRH and assigning a channel to each UE), we formulate the complete resource chain of H-CRAN/C-RAN to further increase the utilization of the cloud-based RANs compared to the previous one-sided approaches. Also, we propose a resource optimization framework which is aware of both the network uncertainties and the different characteristics of C-RAN resource components with respect to the different timescales in scheduling the resources. The former is important especially when utilizing the short-range base stations, while the latter is critical to the new challenges faced in C-RAN.

The remainder of this chapter is organized as follows. Section 3.3 introduces the overview of the problem, network model and essential elements for formulating an SP. In Section 3.4, we formulate the optimal resource allocation problem by means of multi-stage SP. Section 3.5 presents the evaluation results, and finally, Section 3.6 concludes this chapter.

3.3 Problem Description

3.3.1 Problem Overview

In this chapter, we study optimal resource scheduling and sharing for H-CRAN under uncertainty, where the objective is to maximize profit while minimizing power consumption. We assume that there are multiple MNOs operating in the given region, and we focus on MNO$_1$ along with its resources and subscribers. Without loss of generality, we assume there are two MNOs, but it can be easily extended to the case with multiple MNOs. In order to avoid confusion in naming different type of BSs, we use the following notations:

- **MBS** for conventional high-power macro base station operated by MNO$_1$,
- **RRH** for remote radio head operated by MNO$_1$, and
- **AP** for access point[3] operated by MNO$_2$, but they are also accessible to MNO$_1$'s subscribers; in what follows, *base station (BS)* refers to any or all of the above three.

[3] Note that an AP can be RRH, MBS, low-power BS, 802.11-type AP, and so forth depending on the MNO$_2$'s network model.

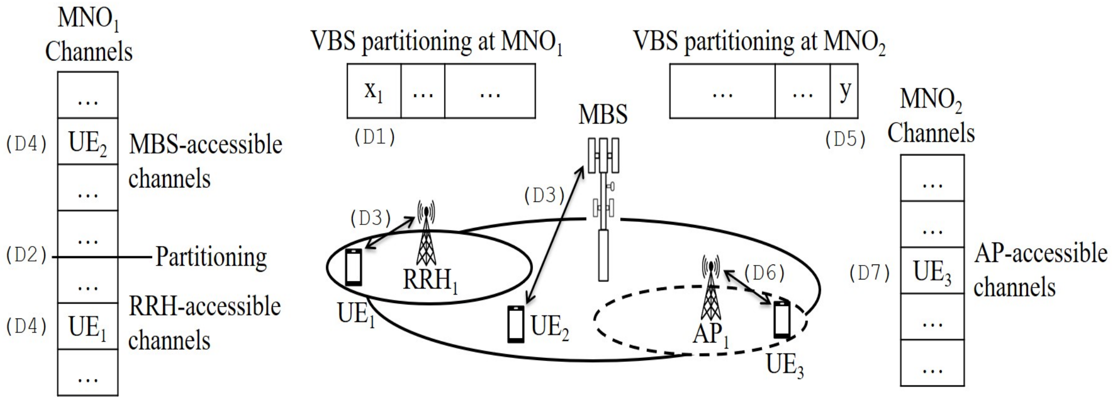

Figure 3.6: Illustration of the proposed resource scheduling method, highlighting the set of decisions to be made. All UEs shown in the figure are MNO$_1$'s subscribers.

In the proposed method, MNO$_1$ makes the following seven decisions from (D1) to (D7) as illustrated in Fig. 3.6. (D1) MNO$_1$ partitions its BBU pool (i.e., the central computing resource), forms VBSs and assigns the VBSs to RRHs. Also, (D2) it divides the available, orthogonal channels into two disjoint sets, one for MBS and the other for RRHs; as a result, there is no cross-tier interference. During operation, (D3) UEs receive service from either MBS or RRH, and (D4) the channel assignment for each UE is determined. In addition, one MNO is allowed to collaborate with another in a way that MNO$_1$'s subscribers can offload their traffic to the other MNO. To do so, (D5) MNO$_1$ has to reserve a certain amount of computing resource at MNO$_2$ in advance. For MNO$_1$, when the aggregate service demand exceeds the network capacity, (D6) it allows UEs to offload their traffic to MNO$_2$'s APs, and at the same time (D7) which channel to use is determined. The aggregate profit to MNO$_1$ is revenue minus the sum of both cost and penalty. The revenue increases in the subscribers' data usage, i.e., number of bits transferred and processed for the subscribers. On the other hand, MNO$_1$ has to pay the cost when it reserves computing resource at MNO$_2$ (i.e., resource reservation fee) and when its subscribers offload their traffic to MNO$_2$'s APs (i.e., offloading cost). Also, penalty is charged to MNO$_1$ whenever there is unmet demand.

In order to effectively responds to the uncertainties of our interest, i.e., users' mobility and their service demand, we formulate the problem in a SP [76] [77] model. SP is a mathematical programming tool in which some of the parameters are uncertain and described by their probability distributions. In contrast to other optimization problems that eliminate uncertainties by assuming constant values for uncertain parameters, SP yields a more robust and practical solution since it takes possible realizations of such parameters into account. As a result, the problem of resource over-/under-provisioning is minimized. Also, the separation of stages and recourse actions [76] in SP result in a structured problem formulation and robust solutions against uncertainties, respectively. In this chapter, *realizations* of an uncertainty parameter refer to the possible values that the parameter can take on, and *scenario* refers to the set of such realizations. One scenario includes multiple realizations, one for each uncertainty parameter, and the set of such scenarios is called scenario tree.

3.3.2 Network Model and Assumptions

On the network is an MBS overlaid by $|\mathcal{M}_1|$ RRHs. The network is operated by MNO$_1$ which has $|\mathcal{U}|$ number of active subscribers. RRHs are sparsely deployed[4] in a planned manner such that there is no intra-tier interference. Each RRH is directly connected to the BBU pool via a wired fronthaul link with limited capacity [78]. The MBS provides coverage, meaning that each UE can always access it; yet, it is not always the case for the wireless links between UEs and RRHs. On the same region, another operator, MNO$_2$, has its own APs installed. Both operators have a limited number of non-overlapping channels each, and thus, there is no inter-network interference. Each UE has two radio interfaces, one for MNO$_1$ and the other for MNO$_2$. Service demand (i.e., downlink traffic rate) of each UE is uniform random with a known mean value.

We assume that MNO$_1$ has an access to historical data (or has a knowledge of known patterns [61]) with respect to UEs mobility. In particular, what we are interested in is the *accessibility* information between UEs and BSs. During operation, a UE periodically listens to the pilot signals from nearby BSs and transmits the list of accessible BSs (i.e., MBS, RRH or AP) in the uplink.

[4] In H-CRAN, the coverage can be provided by MBS while high spectral efficiency can be achieved by RRHs. In fact, a dense deployment of low-power BSs may not be efficient since many of them are often under-utilized [55], and it also increases both interference and energy consumption [60].

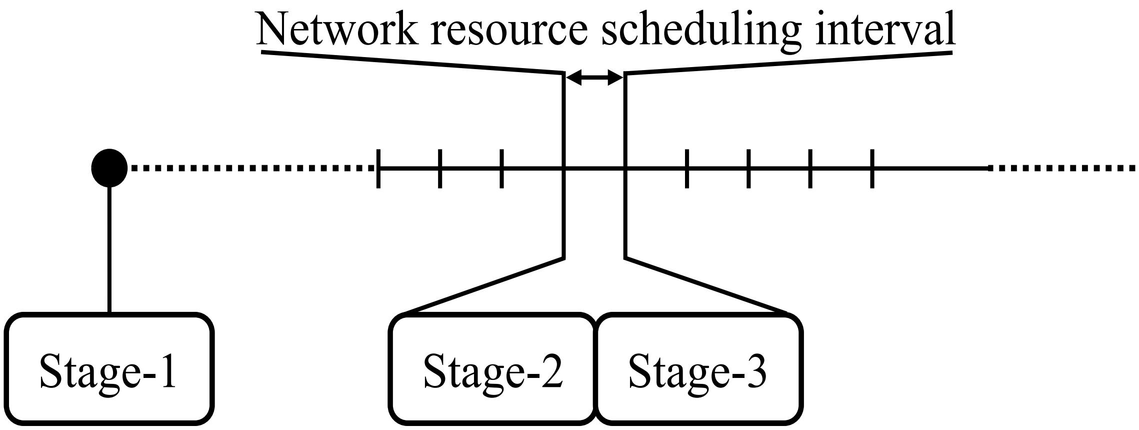

Figure 3.7: In this work, we assume a slotted time frame. The `Stage-1` carries out a long-term resource allocation, while for each time slot, both `Stage-2` and `Stage-3` decisions are made to provide service to users on a small time scale.

After accumulating such information, an MNO can compute the probability that a UE is accessible to a particular BS during a certain period of time, e.g., a day. We also assume an equal per-channel transmission power at a BS, which provides close-to-optimal performance [36][47]. The entire BBU pool is assumed to be continuous and fractional partitioning is allowed. A VBS is assigned to a single RRH, and it cannot be shared among multiple RRHs.

Notations are summarized in Table 3.1, while the others are defined as needed. For indexing purpose, u, i_1, i_2, m_1, m_2 and s are used to indicate a UE, RRH, AP, MNO$_1$'s channel, MNO$_2$'s channel and scenario, respectively. Vectors are in bold and lowercase, e.g., $\mathbf{x}$, $\boldsymbol{\pi}$ and $\mathbf{1}$. Matrices are in uppercase, bold letters, e.g., $\mathbf{A}_s$ and $\mathbf{B}_s$. Uppercase letters in calligraphic font indicate sets e.g., $\mathcal{I}_1$ and $\mathcal{M}_2$.

3.3.3 Decisions and Timeline

Time is partitioned or slotted as shown in Fig. 3.7, where network resource scheduling is carried out at the beginning of each time slot. At the beginning, `Stage-1` is executed, and `Stage-2` and `Stage-3` runs in sequence on each time slot. The first stage, `Stage-1`, is for network planning, `Stage-2` is for service provisioning, and the last stage, `Stage-3`, is for a recourse action. For `Stage-1`, MNO$_1$ makes the following decisions that will last for a while without knowing both the

UEs' locations and their service demand. The BBU pool is partitioned to form VBSs for RRHs, i.e., `(D1)`. Since VBSs are formed by means of VMs, reconfiguring or resizing VMs can cause a significant delay which might be intolerable for cellular networking systems. Channel partitioning, i.e., `(D2)`, also belongs to `Stage-1` since it may cause a significant delay because of the heavy computation load which will be discussed in Section 3.4. Also, reserving MNO_2's computation resource, i.e., `(D5)`, should be done at `Stage-1` since its availability cannot be expected in advance. Such decisions can be made periodically, but with relatively long intervals in-between.

On the other hand, the `Stage-2` and `Stage-3` decisions are repeatedly made on a short time scale, e.g., duration of a resource block in 4G LTE, so that an MNO can satisfy users' QoS as much as possible against the frequently-changing UEs' mobility and their service demand. Once the location and service demand of UEs become known at the beginning of `Stage-2`, the proposed method associates each UE with either an MBS or RRH (if connected), i.e., `(D3)`, and also maps UEs to available channels to provide service, i.e., `(D4)`. If the aggregate service demand at the moment exceeds the network capacity of MNO_1, a recourse action is taken by `Stage-3`. In `Stage-3`, MNO_1 allows UEs with unmet demand to offload their traffic to nearby APs (if connected) with assigning them the best channels for service, i.e., `(D6)` and `(D7)`. If there is no unmet demand at the end of `Stage-2`, no further resource scheduling is needed during `Stage-3`.

3.3.4 Uncertainties and Scenario Representation

The two uncertainties of our interest in this study are UEs' locations and their service demand. The locations are represented by a more abstract term, which is accessibility between UEs and RRHs/APs. The corresponding accessibility information are denoted by $\mathbf{V}$ (for UEs-RRHs) and $\mathbf{W}$ (for UEs-APs). Both $\mathbf{V}$ and $\mathbf{W}$ are uncertain parameters whose distributions are assumed to be known from historical data. Both $\mathbf{V}$ and $\mathbf{W}$ are represented by matrices of 0's and 1's, and determined by UE's location and the channel quality. Please note that a UE can always access to MBS by assumption. Users' service demand is denoted by $\mathbf{r}$ whose distribution follows a uniform random with known mean values. Let $\xi = (\mathbf{V}, \mathbf{W}, \mathbf{r})$ be a *snapshot* of the network describing the users' accessability information and service demand. The ξ also is a uncertain parameter; to be specific, it is a set of uncertain parameters.

SP handles such uncertain parameters by means of scenarios. In a particular scenario s, the network snapshot is known to be $\xi_s = (\mathbf{V}_s, \mathbf{W}_s, \mathbf{r}_s)$. In scenario s, the accessibility between UEs and RRHs, the accessibility between UEs and APs, and the UEs' service demand, respectively, are known to be $\mathbf{V}_s$, $\mathbf{W}_s$, $\mathbf{r}_s$, and they are drawn from their corresponding distributions. A scenario is a realized snapshot of the network, and it can occur at any certain time instance. The set of such scenarios constitutes a scenario tree. A complete scenario tree, $\mathcal{T}$, is a Cartesian product of all possible realizations of all uncertain parameters, describing all possible snapshots that can occur at any time instance. Since the continuous random variable $\mathbf{r}$ has an infinite number of realizations, making an infinite number of scenarios, we use a discretization method proposed by [79] to reduce the problem size, and thus, to be able to find the optimal solution in a finite time; it will be discussed shortly in both Section. 3.4.3 and Section. 3.5.2.

3.4 Problem Formulation

Among the decisions that are jointly considered in this chapter, the coupling of channel partitioning and channel allocation is what makes the whole problem to be non-convex and intractable. As a reminder, channel partitioning is the partition of available channels into two sets, one for MBS and the other for RRHs, while channel allocation is to assign users channels to provide service. In order to make the problem tractable, we propose an iterative optimization algorithm such that channel partitioning is separately handled from the rest decisions.

3.4.1 Decomposition and Iterative Search

The proposed iterative resource allocation algorithm is given in Algorithm 2. Both ρ^* and ρ^t indicate the optimal objective value and the objective value on t-th iteration, respectively. The composite variable Ω is a set of the decision variables we consider in the proposed problem formulation, which will be introduced shortly. Both Ω^* and Ω^t, respectively, denote the optimal

Algorithm 2 Iterative resource allocation algorithm

```
1:  $\rho^* \leftarrow -\infty$                                   // optimal objective value
2:  $\Omega^* \leftarrow$ Null                                   // optimal solution set
3:  $\boldsymbol{\pi}^* \leftarrow$ Null                         // optimal channel partition
4:  $\boldsymbol{\pi} \leftarrow \mathbf{0}_{|\mathcal{M}_1|}$    // initialization
5:  for $t \leftarrow 1$ to $|\mathcal{M}_1| - 1$ do
6:      $\boldsymbol{\pi}[t] \leftarrow 1$                       // set $t$-th element to 1
7:      Get $(\rho^t, \Omega^t)$ by solving DEP (P. 3.4)
8:      if $\rho^t > \rho^*$ then
9:          $\rho^* \leftarrow \rho^t$, $\Omega^* \leftarrow \Omega^t$, $\boldsymbol{\pi}^* \leftarrow \boldsymbol{\pi}$
10:     end if
11: end for
12: Return $(\Omega^*, \boldsymbol{\pi}^*)$
```

decision and the decision found on t-th iteration. Lines 1–3 initialize the optimal objective value, optimal solution set and optimal channel partitioning decision in sequence. Line 4 initializes the channel partitioning vector $\boldsymbol{\pi}$ to a zero vector. On each iteration of the **for** loop (lines 5–11), the number of channels allocated to MBS is increased from 1 to $|\mathcal{M}_1| - 1$ by 1 (line 6), where $\mathcal{M}_1$ is the index set of channels accessible to MNO$_1$.

The $\boldsymbol{\pi} \in \{0,1\}^{|\mathcal{M}_1|}$ is an MNO$_1$'s channel allocation vector for MBS, where having 1 or 0 at t-th element indicates that the corresponding channel t is accessible to MNO$_1$'s MBS or not, respectively. We assume consecutive channel allocation, i.e., the set of channels accessible to MBS is from 1 to m_1 whereas that to SBS is from $m_1 + 1$ to $|\mathcal{M}_1|$. Also, there should be at least one channel available to MBS and RRHs each. The accessible channels for MBS and RRH do not overlap, and thus there is no cross-tier interference. RRHs operate on the channels that are marked by zeros in $\boldsymbol{\pi}$. We denote such RRH-accessible channels by $\overline{\boldsymbol{\pi}} = \mathbf{1}_{|\mathcal{M}_1|} - \boldsymbol{\pi}$, where $\mathbf{1}_{|\mathcal{M}_1|}$ is a vector of all 1's with $|\mathcal{M}_1|$ entries therein. In $\overline{\boldsymbol{\pi}}$, RRH-accessible channels are marked by 1's. As a result, given the power budget P^M_{max} for an MBS and $= P^R_{max}$ for an RRH, the per-channel power use of MBS and RRH becomes $P^M_{max}/|\boldsymbol{\pi}|$ and $P^R_{max}/|\overline{\boldsymbol{\pi}}|$, respectively.

In line 7, Algorithm 1 solves DEP (deterministic equivalent problem), which will be introduced in Section. 3.4.3. In a nutshell, we first form an energy-aware profit maximization problem in

SP, and then, transform it into a DEP instance so that we can solve it by the proposed iterative algorithm with a computer solver, e.g., Matlab [49]. Lines 8–9 compare the best objective value so far to the current one, and record the current result if it has a higher objective value. Once the algorithm terminates, it returns both Ω^* and $\boldsymbol{\pi}^*$ (line 12), with which MNO_1 can configure its network.

Please note that Algo. 2 may not produce a global optimal solution if the optimal solution can only be found with a non-consecutive channel partitioning. However, the assumption on consecutive channel allocation helps to reduce the complexity of the entire algorithm significantly. In addition, if the globally optimal solution exists for consecutive channel partitioning, the proposed algorithm can produce the same optimal solution as global optimum.

3.4.2 Multi-Stage Stochastic Programming

`Stage-1` is for network planning, where the long-term decisions are made without knowing the actual location and demand of UEs. MNO_1 assigns a fraction of BBU pool to each RRH after forming each segment of BBU pool into a VBS (or VM). Also, MNO_1 reserves a certain amount of processing resource at MNO_2's network in preparation for a sudden increase of service demand. The `Stage-1` problem is shown below (called P. 3.1).

$$\max_{\mathbf{x},y}. \quad -\delta \sum_{i_1 \in \mathcal{I}_1} x_{i_1} - \epsilon \cdot y + \mathcal{Q}(\mathbf{x}, y) \tag{3.1a}$$

$$\text{subject to} \quad \sum_{i_1 \in \mathcal{I}_1} x_{i_1} \leq 1, \tag{3.1b}$$

$$\forall i_1 : \; 0 \leq x_{i_1} \leq 1, \tag{3.1c}$$

$$0 \leq y \leq 1, \tag{3.1d}$$

$$\forall i_1 : \; x_{i_1} \cdot K_1 \leq c_{i_1}, \tag{3.1e}$$

The objective (Eq. 3.1a) is to maximize the total profit, i.e., revenue minus cost. The first term is the OPEX for operating the BBU pool, where $x_{i_1} \in \mathbf{x}$ is the fraction of BBU allocated to RRH $i_1 \in \mathcal{I}_1$, and δ is a conversion parameter from x_{i_1} to expense ($). The second term is the rental fee, where y is the fraction of computing resource that MNO$_1$ reserves at MNO$_2$ in advance, and ϵ is a conversion parameter from y to the rental fee (\$). The third term is the expected profit from Stage-2, i.e., $\mathcal{Q}(\mathbf{x}, y) = \mathbb{E}_\xi[Q_s(\mathbf{x}, y; \xi_s)]$, where $\xi = (\mathbf{V}, \mathbf{W}, \mathbf{r})$ is the uncertain parameter describing a network snapshot (i.e., accessibility and demand), $\xi_s = (\mathbf{V}_s, \mathbf{W}_s, \mathbf{r}_s)$ is a realized scenario indexed by s, and $Q_s(\mathbf{x}, y; \xi_s)$ is the Stage-2 objective value given the Stage-1 decisions (i.e., $\mathbf{x}$ and y) and a realized scenario of ξ_s.

For the notation $Q_s(\mathbf{x}, y; \xi_s)$, what comes before and after the semicolon, respectively, is the list of decisions made at the previous stage and the realized scenario (i.e., realizations of uncertain parameters that become known at the stage). For example, ξ_s refers to a particular scenario that has been taken at the beginning of Stage-2. For scenario s, $\mathbf{V}_s$ and $\mathbf{W}_s$ are the realizations denoting the accessibility between UEs and RRHs and UEs and APs, respectively, and $\mathbf{r}_s$ is a realization of UEs' service demand in scenario s. The aggregate BBU resource usage should not exceed the limit of 1 (Eq. 3.1b). Each RRH can be assigned a fraction of BBU (Eq. 3.1c). The portion of the computation resource to borrow from MNO$_2$ should be non-negative, and it cannot exceed the limit of 1 (Eq. 3.1d). The network is fronthaul-constrained (Eq. 3.1e), where c_{i_1} is the fronthaul capacity for the link connected to RRH i_1. The constant K_1 is a conversion parameter from a fraction of BBU to the actual amount of bits that can be processed in unit time, i.e., conversion from [0,1] to bps.

Stage-2 is for service provisioning, whose problem formulation is shown below (called P. 3.2). Given the decisions made in Stage-1 as well as the scenario that has become known, MNO$_1$ schedules its own networking resources and provides service to its subscribers. Please note that out of three realizations, i.e., $\mathbf{V}_s$, $\mathbf{W}_s$ and $\mathbf{r}_s$, that become known in Stage-2, MNO$_1$ makes use of only both $\mathbf{V}_s$ and $\mathbf{W}_s$. As a reminder, MNO$_1$ utilizes its own networking resource in Stage-2. The $\mathbf{W}_s$ will be used in Stage-3 if MNO$_1$ has to offload some of the users' demand to MNO$_2$.

$$Q_s(\mathbf{x}, y; \xi_s) :=$$

$$\max_{\mathbf{A}_s}. \ \alpha(\sum_{u \in \mathcal{U}} \sum_{m_1 \in \mathcal{M}_1} a^u_{0,m_1,s} \cdot f^u_{0,m_1})$$

$$- \theta \cdot \tau^M \sum_{u \in \mathcal{U}} \sum_{m_1 \in \mathcal{M}_1} a^u_{0,m_1,s}$$

$$+ \alpha(\sum_{i_1 \in \mathcal{I}_1} \sum_{u \in \mathcal{U}} \sum_{m_1 \in \mathcal{M}_1} a^u_{i_1,m_1,s} \cdot f^u_{i_1,m_1})$$

$$- \theta \cdot \tau^R \sum_{i_1 \in \mathcal{I}_1} \sum_{u \in \mathcal{U}} \sum_{m_1 \in \mathcal{M}_1} a^u_{i_1,m_1,s}$$

$$+ H_s(y, \mathbf{A}_s; \xi_s) \tag{3.2a}$$

subject to

$$\forall u : \sum_{m_1 \in \mathcal{M}_1} a^u_{0,m_1,s} + \sum_{i_1 \in \mathcal{I}_1} \sum_{m_1 \in \mathcal{M}_1} a^u_{i_1,m_1,s} \leq 1, \tag{3.2b}$$

$$\forall u, m_1 : 0 \leq a^u_{0,m_1,s} \leq \boldsymbol{\pi}_{m_1}, \tag{3.2c}$$

$$\forall u, m_1, i_1 : 0 \leq a^u_{i_1,m_1,s} \leq v^u_{i_1,s} \cdot \overline{\boldsymbol{\pi}}_{m_1}, \tag{3.2d}$$

$$\forall i_1 : \sum_{u \in \mathcal{U}} \sum_{m_1 \in \mathcal{M}_1} a^u_{i_1,m_1,s} \leq o_{i_1}, \tag{3.2e}$$

$$\forall i_1 : \sum_{u \in \mathcal{U}} \sum_{m_1 \in \mathcal{M}_1} a^u_{i_1,m_1,s} \cdot f^u_{i_1,m_1} \leq x_{i_1} K_1, \tag{3.2f}$$

$$\forall m_1 : \sum_{u \in \mathcal{U}} a^u_{0,m_1,s} \leq \boldsymbol{\pi}_{m_1}, \tag{3.2g}$$

$$\forall i_1, m_1 : \sum_{u \in \mathcal{U}} a^u_{i_1,m_1,s} \leq \overline{\boldsymbol{\pi}}_{m_1}, \tag{3.2h}$$

$$\forall u : \sum_{m_1 \in \mathcal{M}_1} a^u_{0,m_1,s} \cdot f^u_{0,m_1}$$

$$+ \sum_{i_1}^{N_{RRH}} \sum_{m_1 \in \mathcal{M}_1} a^u_{i_1,m_1,s} \cdot f^u_{i_1,m_1} \leq r^u_s. \tag{3.2i}$$

In P. 3.2, the objective (Eq. 3.2a) is to maximize profit. The profit increases in the number of bits carried in downlink for subscribers, but decreases in the power consumption. The constant α is the conversion parameter from the number of serviced bits to what users has to pay (i.e., profit to be given to MNO$_1$), $a^u_{i_1,m_1,s} \in [0, 1]$ is the fraction of time [36] [80] [81] that UE u makes an association with an MBS (if $i_1 = 0$) or RRH (if $i_1 \neq 0$) and receives service on channel m_1, $f^u_{i_1,m_1}$ is the traffic rate (or bandwidth) at which UE u is receiving data from an MBS (if $i_1 = 0$) or RRH (if $i_1 \neq 0$) on channel m_1, θ is the conversion parameter from the amount of power consumed to the charge for the usage, τ^M (or τ^R) is the per-channel power use of MBS (or RRH), and $H_s(y, \mathbf{A}_s; \xi_s)$ is the profit from the `Stage-3` program. `Stage-3` is for an recourse action and is deterministic once a scenario (or network snapshot) becomes known at the beginning of `Stage-2`. Thus, $H_s(\cdot)$ is not an expected profit. Also, the `Stage-3` is launched only when there is unmet demand in `Stage-2`.

Each UE is charged for its data usage by the first and the third term on the objective function depending on which type of BS it is associated with. The term $f^u_{i_1,m_1}$ is the traffic rate, bandwidth or channel capacity, defined as:

$$f^u_{i_1,m_1} = \begin{cases} \Delta \log_2(1 + \frac{\Psi^u_{i_1,m_1} \cdot \tau^M}{\Delta \sigma^2}), & \text{if } i_1 = 0 \text{ (i.e., MBS).} \\ \Delta \log_2(1 + \frac{\Psi^u_{i_1,m_1} \cdot \tau^R}{\Delta \sigma^2}), & \text{otherwise (i.e., RRH).} \end{cases}$$

where Δ is the channel bandwidth, $\Psi^u_{i_1,m_1}$ is the channel gain between UE u and MBS/RRH i_1 over the channel m_1 and σ^2 is the per-Hz noise power. Also, the second and fourth term in the objective function charges MNO_1 for power consumption.

For each UE, the aggregate amount of time to use MNO_1's BSs cannot exceed a unit time (Eq. 3.2b). The constraints Eq. 3.2c and Eq. 3.2d indicate that each UE can access MBS and RRH, respectively, only through the channels accessible to the corresponding BS. Note that $v^u_{i_1,s} \in \mathbf{V}_s$ is a binary indicator, telling whether UE u is within the coverage of RRH i_1 or not, denoted by 1 or 0, respectively. However, we do not need it in Eq. 3.2c which is for MBS, since an MBS is assumed to be always accessible to provide coverage. The channel allocation vector for MBS and RRH are π and $\overline{\pi}$, respectively. The number of UEs that an RRH can handle at the same time is limited by o_{i_1} (Eq. 3.2e), while that of an MBS is assumed to be large. The fronthaul link between each RRH and the BBU pool has a limited capacity, meaning that the aggregate amount of downlink traffic carried over the link is limited (Eq. 3.2f). The constraints in Eq. 3.2g and Eq. 3.2h are to guarantee that an MBS and RRH, respectively, cannot use a channel for more than a unit time, if accessible. Lastly, the total number of bits that a UE receives from MBS and RRHs does not exceed its demand (Eq. 3.2i).

`Stage-3` is for a recourse action. For those UEs whose service demand is not fully satisfied, they are allowed to use MNO_2's resource through APs by offloading their traffic to MNO_2. This is the case when the entire or a part of the MNO$_1$'s network is saturated, and thus MNO$_1$ cannot meet the demand of all UEs. The problem formulation is given below (called P. 3.3).

$$
\begin{aligned}
H_s(y, \mathbf{A}_s;\xi_s) :=& \\
\max_{\mathbf{B}_s,\widetilde{\mathbf{r}}_s}. \ & \alpha(\sum_{i_2\in\mathcal{I}_2}\sum_{u\in\mathcal{U}}\sum_{m_2\in\mathcal{M}_2} b^u_{i_2,m_2,s}\cdot g^u_{i_2,m_2}) \\
& -\beta(\sum_{i_2\in\mathcal{I}_2}\sum_{u\in\mathcal{U}}\sum_{m_2\in\mathcal{M}_2} b^u_{i_2,m_2,s}\cdot g^u_{i_2,m_2}) \\
& -\gamma\sum_{u\in\mathcal{U}}\widetilde{r^u_s} && (3.3a)
\end{aligned}
$$

subject to

$$\forall u: \ \sum_{i_2\in\mathcal{I}_2}\sum_{m_2\in\mathcal{M}_2} b^u_{i_2,m_2,s}\leq 1, \tag{3.3b}$$

$$\forall u, i_2, m_2: \ 0\leq b^u_{i_2,m_2,s}\leq w^u_{i_2,s}, \tag{3.3c}$$

$$\forall i_2: \ \sum_{u\in\mathcal{U}}\sum_{m_2\in\mathcal{M}_2} b^u_{i_2,m_2,s}\leq d_{i_2}, \tag{3.3d}$$

$$\forall i_2: \ \sum_{u\in\mathcal{U}}\sum_{m_2\in\mathcal{M}_2} b^u_{i_2,m_2,s}\cdot g^u_{i_2,m_2}\leq z_{i_2}, \tag{3.3e}$$

$$\forall u, i_2: \ \sum_{m_2\in\mathcal{M}_2} b^u_{i_2,m_2,s}\leq 1, \tag{3.3f}$$

$$
\begin{aligned}
\forall u: \ & \sum_{m_1\in\mathcal{M}_2} a^u_{0,m_1,s}\cdot f^u_{0,m_1} \\
& +\sum_{i_1\in\mathcal{I}_1}\sum_{m_1\in\mathcal{M}_1} a^u_{i_1,m_1,s}\cdot f^u_{i_1,m_1} \\
& +\sum_{i_2\in\mathcal{I}_2}\sum_{m_2\in\mathcal{M}_2} b^u_{i_2,m_2,s}\cdot g^u_{i_2,m_2} \\
& +\widetilde{r^u_s} = r^u_s && (3.3g)
\end{aligned}
$$

$$\forall u: \ \widetilde{r^u_s}\geq 0, \tag{3.3h}$$

$$
\begin{aligned}
& \sum_{i_2\in\mathcal{M}_2}\sum_{u\in\mathcal{U}}\sum_{m_2\in\mathcal{M}_2} b^u_{i_2,m_2,s}\cdot g^u_{i_2,m_2} \\
& \leq y\cdot K_2. && (3.3i)
\end{aligned}
$$

`Stage-3` has a similar problem structure to `Stage-2`. However, in this stage, MNO$_1$ lets its subscribers use MNO$_2$'s networking resource in order to prevent service outage as much as possible; otherwise, MNO$_1$ has to pay the penalty for unmet demand. To do so, MNO$_1$ has to pay the necessary costs in both `Stage-1` and `Stage-3`. The cost that MNO$_1$ is charged at `Stage-1` is for reserving MNO$_2$'s resource (i.e., reservation fee), while the cost at `Stage-3` is for the actual usage of it (i.e., offloading cost).

The objective (Eq. 3.3a) is to maximize the profit, which is a function of the amount of both the serviced bits via MNO$_2$ and the unmet demand. The parameter β is an offloading cost to MNO$_2$ (i.e., cost for using MNO$_2$'s networking resource) and γ is a penalty for unmet demand. Given the two parameters, the three terms in Eq. 3.3a correspond to the profit from serviced bits, the cost for traffic offloading and the penalty for the unmet demand in sequence. The total amount of time to use MNO$_2$'s resource cannot exceed the limit of 1 (Eq. 3.3b), where $b^u_{i_2,m_2,s} \in \mathbf{B}_s$ is a fraction, [0,1] of time UE u associates with AP i_2 and receives service on channel m_2 in scenario s. A UE can access an AP only when it is within the coverage of the AP (Eq. 3.3c), i.e., when $w^u_{i_2,s} = 1$. Each AP has a maximum number of UEs to which it can provide service simultaneously (Eq. 3.3d) with a limited aggregate bandwidth capacity (Eq. 3.3e). Each channel cannot be occupied/used for more than a unit time (Eq. 3.3f).

For each UE, the total received service and unmet demand should equal its service demand (Eq. 3.3g). In the `Stage-3` formulation, the service demand requirement is expressed by an equality constraint by using a slack variable $\widetilde{r^u_s}$ as in Eq. 3.3g. The amount of unmet demand for each UE is denoted by $\widetilde{r^u_s}$ which is non-negative (Eq. 3.3h). The Eq. 3.3i indicates that the aggregate amount of traffic offloaded to MNO$_2$ should not exceed the limit determined by the decision made in `Stage-1`. The conversion parameter K_2 is multiplied to y which is the fraction of the computing resource reserved at MNO$_2$, resulting in the actual amount of offloaded bits that can be processed by MNO$_2$ in unit time.

The Fig. 3.8 summarizes the profit and cost model proposed in this study, which is one of the major contributions. Depending on the decisions to make in each stage, MNO$_1$ pays fee/cost and/or

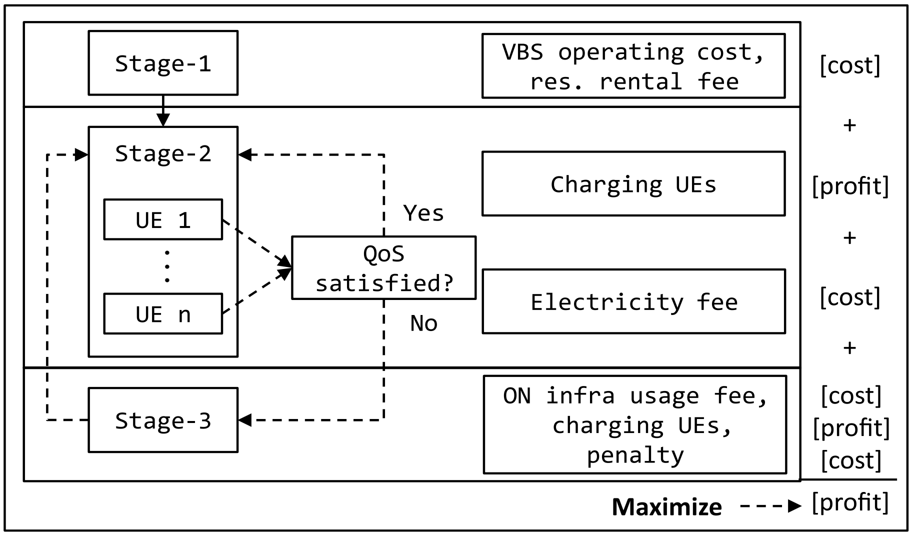

Figure 3.8: The proposed profit and cost model.

makes a profit. In `Stage-1`, the MNO_1 configures its networking resources without providing any service to users, and thus, there is no profit to make. In `Stage-2` and `Stage-3`, on the other hand, users receive data service from MBS/RRH or AP, respectively, and from which MNO_1 makes profit; i.e., users are charged for the amount of data received. The MNO_1 pays for the power consumed to provide data service to users in `Stage-2`. Besides, upon the use of the MNO_2's networking resource to offload the users' traffic, MNO_1 pays for it to compensate for the MNO_2's operational cost in `Stage-3`. Lastly, if there is any unmet demand, MNO_1 pays a penalty in proportion to the total amount in `Stage-3`.

3.4.3 Deterministic Equivalent Problem (DEP)

The last step is to transform the SP model into a computable deterministic equivalent, DEP, to be able to solve it with a computer solver. To begin with, given the distribution of the uncertain parameters, we can find the probability p_s of each scenario s in the scenario tree $\mathcal{T}$. By taking the expectation of the objective values of the stages that are dependent upon the realization of uncertainty parameters therein, we get the deterministic formulation of the SP as below (called P. 3.4).

$$\begin{aligned}\max_{\substack{\mathbf{x},y,\\ \mathbf{A},\\ \mathbf{B},\tilde{\mathbf{r}}}}. \quad & -\delta \sum_{i_1\in\mathcal{I}_1} x_{i_1} - \epsilon\cdot y \\ & + \sum_{s\in\mathcal{S}} p_s[Q_s(\mathbf{x},y;\xi_s) + H_s(y,\mathbf{A}_s;\xi_s)] \qquad (3.4a)\\ \text{s.t.}\ & \text{constraints in (3.1b)-(3.1e), (3.2b)-(3.2i), (3.3b)-(3.3i),}\end{aligned}$$

where *s.t.* is short for subject to.

The objective (Eq. 3.4a) is the sum of the `Stage-1` objective value and the expectation of the remaining stages' objectives. Constraints from all stages are then followed to make the DEP equivalent to the SP formulation. The Algo. 2 in Section. 3.4.1 solves DEP (P. 3.4) on each iteration as shown in line 7. As aforementioned in 3.3.3, we can apply a discretization method to the continuous uncertainty parameter, i.e., the service demand $\mathbf{r}$ which follows the continuous uniform distribution, to let it have a finite support; please refer to Appendix A for details. As a result, the set of scenarios in $\mathcal{T}$ also has a finite support, meaning that we can find the probability p_s of each scenario ξ_s in the complete scenario tree $\mathcal{T}$. Here, p_s is the probability that ξ takes on ξ_s and $\mathcal{S}$ is the index set of scenarios, where $|\mathcal{S}| = |\mathcal{T}|$.

3.5 Evaluation

3.5.1 Network Configuration

MBS has a coverage radius of 300 meters, within which four RRHs are located. Each RRH has a radius of 100 meters. RRHs have a keep-away distance of 100 meters from the MBS tower, and RRHs are equally spaced. APs are located in a similar manner as RRHs, but the location of an AP is not within any of the RRHs on the network. The coverage radius of an AP is 150 meters. For ϵ and δ, we assume that both MNO_1 and MNO_2 are using cloud computing service for processing user data, such as Amazon EC2 [82]. In particular, we have chosen the reserved instance pricing model of the general purpose instance, m4. The penalty for unmet demand is triple the charge for data use. Service demand of each user follows uniform distribution whose mean value is randomly drawn from $[0, 20]$ Mbps, and the actual service demand of each UE changes over time.

We have used the channel model and parameters in [42]. The distance-dependant path loss from MBS to UE and RRH to UE is PL (dB) $= 128.1 + 37.6 \cdot \log_{10}(R)$ and PL (dB) $= 140.7 + 36.7 \cdot \log_{10}(R)$, respectively, where R is the distance in km. Log-normal random variable with standard deviation of 10 dB is used to model shadowing, and a normalized Rayleigh for small-scale fading effect. The channel bandwidth is 1.25 MHz, and the total transmission power for MBS and RRH is 20 W (43 dBm) and 100 mW (20 dBm), respectively. The noise power density is -174 dBm/Hz [73]. UEs with mobility are initially randomly located, and from the location of UEs, the accessibility between UE and RRH as well as UE and AP are determined periodically.

3.5.2 Solution Procedure

Given that the possible realizations the users' service demand $\mathbf{r}$ are infinite, the DEP problem (P. 3.4) is intractable due to the explosive data size. Thus, we apply the discretization method to $\mathbf{r}$ as aforementioned in Section. 3.3.4. In particular, we approximate each UE's demand into three discrete values to indicate low, mean and high demand while minimizing the errors between the the original distribution (i.e., continuous uniform random) and the 3-point approximation for each user; please refer to Appendix for detail. Due to the the vast size of the data set, however, it is still not efficient or even possible to solve the problem with complete scenario tree[5]. Thus, we apply the sample average approximation (SAA) method [76] to reduce the problem size (i.e., the number of scenarios to be solved). That is, instead of solving the DEP (P. 3.4) for the complete scenario tree $\mathcal{T}$, we construct a reduced scenario tree $\mathcal{T}'$ by randomly sampling a subset of scenarios from $\mathcal{T}$. Let $\mathcal{S}'$ be the scenario index set corresponding to $\mathcal{T}'$. We first replace the expected `Stage-2` profit in P. 3.1 by a Monte Carlo estimate $\mathcal{Q}(\mathbf{x}, y)' = \sum_{s\in\mathcal{S}'} Q_s(\mathbf{x}, y; \xi_s)/|\mathcal{S}'|$, where each scenario in $\mathcal{T}'$ is equally likely and $\xi_s \in \mathcal{T}'$. Next, we replace the objective function of DEP (P. 3.4) by what

[5] Note that that the cardinality of the complete scenario tree is $(N_L \cdot N_Q)^{N_{UE}}$, where N_L and N_D are the numbers of realizations of the uncertainties related to the UEs' location and their service demand, respectively, and N_{UE} is the number of UEs.

follows:

$$\max_{\substack{\mathbf{x},y,\\ \mathbf{A},\\ \mathbf{B},\tilde{\mathbf{r}}}}. \quad -\delta \sum_{i_1\in\mathcal{I}_1} x_{i_1} - \epsilon\cdot y \\ + \frac{1}{|\mathcal{S}'|}\sum_{s\in\mathcal{S}'}[Q_s(\mathbf{x},y;\xi_s)+H_s(y,\mathbf{A}_s;\xi_s)]$$

Due to the use of reduced, randomly sampled scenarios in the SAA method, optimal solutions vary with respect to which scenarios have been chosen to construct a *reduced* scenario tree $\mathcal{T}'$. In this regard, we will show both stability and quality of solution at the end of this section. We have implemented and evaluated the proposed method as well as the ones for comparison on Matlab and CVX [50]. We have constructed a scenario tree with 30 scenarios chosen at uniform random for Sections 3.5.3, 3.5.4 and 3.5.5. In Section 3.5.6, we construct three more scenario trees in the same manner in order to study the solution stability and quality. For each scenario tree we have simulated the network for 1,000 seconds.

3.5.3 Effect of Uncertainties

First, we have studied the effect of different ways of handling uncertainties on performance. We have compared the proposed method to three different methods that completely or partially ignore the uncertainties. To be specific, the four algorithms to be compared with each other are as follows.

- **CMCD** (constant mobility, constant demand) which assumes constant UEs' mobility and constant service demand[6],
- **CM** (constant mobility) which assumes constant UEs' mobility, but considers uncertain demand,
- **CD** (constant demand) which assumes constant demand, but considers uncertain UEs' mobility, and
- **proposed** which considers the uncertainties in both mobility and demand.

[6] In the context of SP, this type of problem is called expected value (EV) problem.

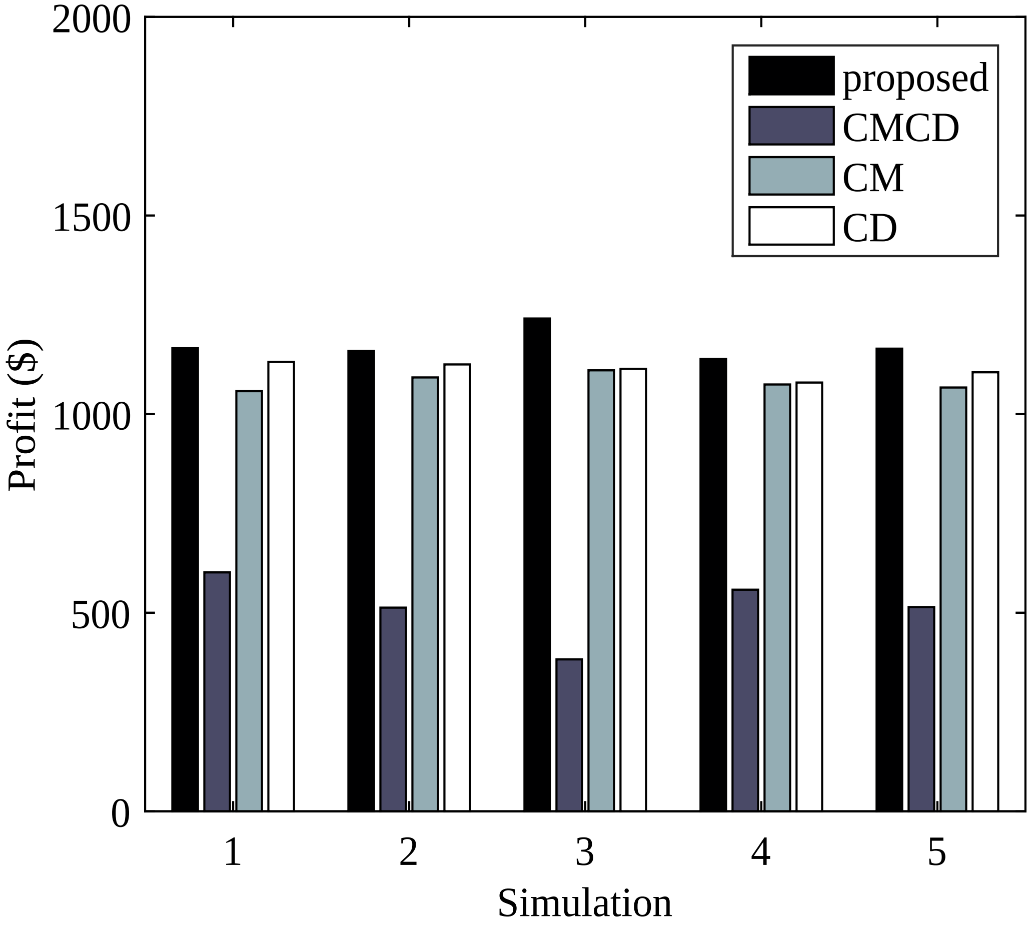

Figure 3.9: Comparison of the optimal profit achieved with respect to different ways of handling uncertainties over five runs of simulation.

Fig. 3.9 shows the optimal profit achieved by the four methods for five runs of simulation. Clearly, CMCD yields the least profit. This is because CMCD ignores the uncertainties in both mobility and service demand by assuming constant values for all uncertain parameters. Thus, CMCD suffers from the under-provisioning problem; i.e., the reserved resource is not enough to handle the actual demand. An insufficient amount of network resource reserved by CMCD soon results in the network saturation, causing users to experience service outage to a large degree. On the other hand, the proposed method, CM and CD consider uncertainties present on the network. Such an awareness makes those methods reserve more resource than CMCD, and thus, the three methods are likely to suffer less from the under-provisioning problem. Among the three that partially or completely consider the uncertainties, the proposed method achieves the highest profit, implying the importance of the comprehensive consideration of the unpredictability. In other words, since CM and CD are unaware of the uncertainty in mobility and service demand, respectively, they have a narrower view of what can happen on the network than the proposed method has.

One interesting finding from comparing CM to CD in Fig. 3.9 is that over all simulations, the performance of CD is no less than that of CM. From this observation, we can conclude that the uncertainties in mobility has a larger effect on the profit than the varying service demand. When there are short-range BSs are deployed and used, the connectivity between such BSs and UEs are important because the short-range BSs can provide high throughput to users with much low power compared to MBSs by taking advantage of the short distance to users. However, the availability of such BSs to users varies significantly over time due to the short service range.

The Fig. 3.10 depicts the average amount of unmet demand per user over five runs of simulation. It also implies whether each method suffers from the resource under-provisioning problem or not. CMCD does not consider any of the network uncertainties, and thus, it optimizes the networking resource only based on the constant parameters for users' mobility and service demand. Due to this shortsighted view over the network, CMCD reserves the least amount of the networking resource and fulfills the constant service demand from the users located at fixed locations. As a result, CMCD suffers most from the service outage. On the other hand, the rest three schedule more resource in order for them to be ready for many different scenarios that are likely to occur.

Although the three methods consider the uncertainties, all of them have non-zero unmet demand. This shows a tradeoff between the penalty for unmet demand and the cost to pay for reserving and utilizing the networking resource. Reserving too much resource on BBU pool and also from MNO$_2$ may result in a zero unmet demand at all times. However, such an over-provisioning approach can leave a significant portion of the resource unused in most of the time, making the solution less cost-effective. In this regard, what Fig. 3.10 shows is that instead of having all demand perfectly satisfied, allowing a negligible amount of QoS degradation gives a higher profit gain

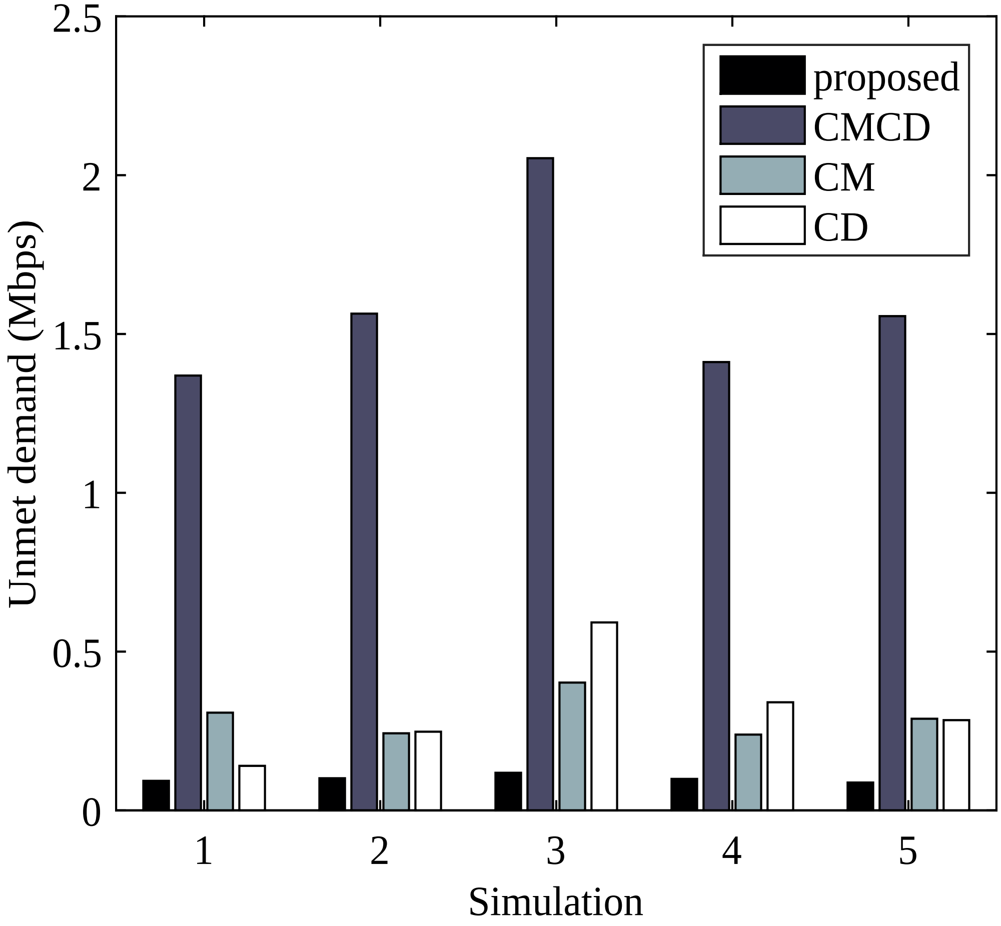

Figure 3.10: Comparison of the average unmet demand per user with respect to different ways of handling uncertainties over five runs of simulation.

overall. Again, the proposed method results in the least amount of unmet demand among the four, showing that it is beneficial to be aware of uncertainties as much as possible.

Continuing the discussion on the mean unmet demand in the previous figure, Fig. 3.11 shows why are the methods that ignore uncertainties suffering from such a large unmet demand. The y axis in Fig. 3.11 indicates how much resource does MNO_1 borrow from MNO_2. In the case of the proposed method, having more knowledge on uncertainties let MNO_1 borrow more resource from MNO_2. However, that is not the case to CMCD. Since CMCD is informed only of one snapshot of

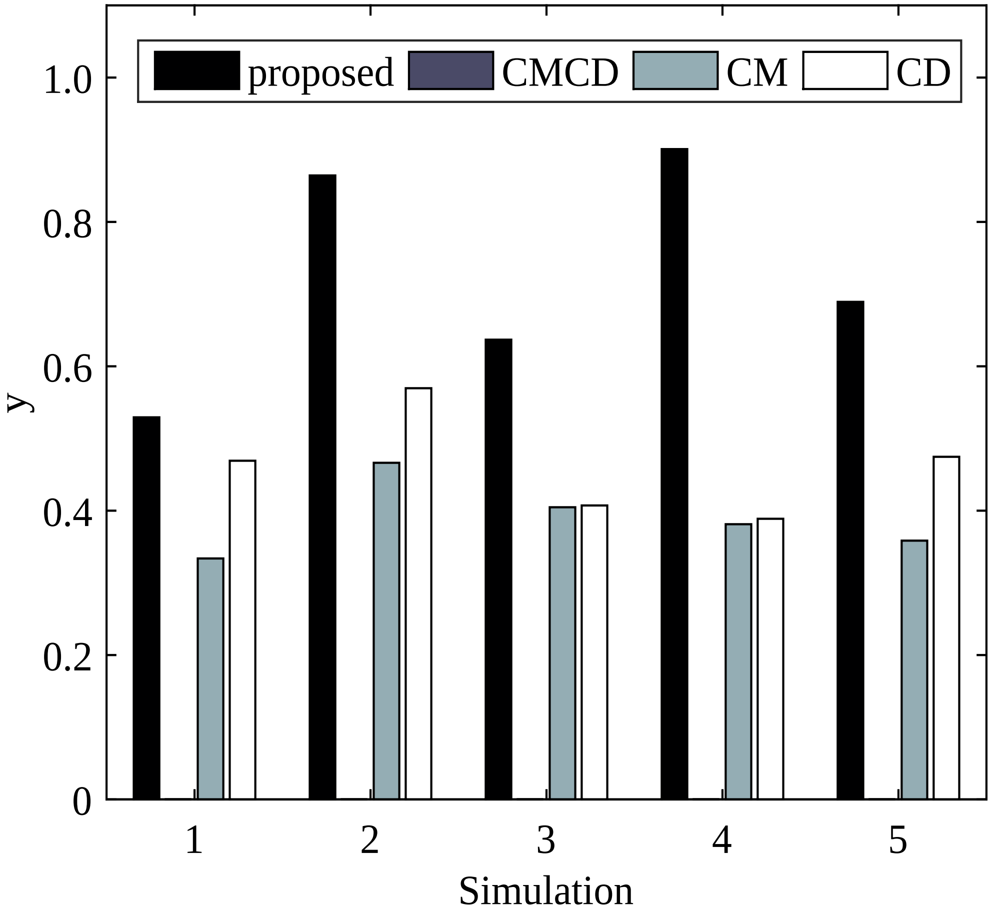

Figure 3.11: Comparison of the fraction (y) of the resource that MNO_1 borrows from MNO_2 at `Stage-1` with respect to different ways of handling uncertainties over five runs of simulation.

the networks as to users' location and service demand, CMCD does not borrow any resource from MNO_2, resulting in a large unmet demand in practice as it can be seen in the previous Fig. 3.10.

Next, we have measured the energy consumption and energy efficiency performance. First, the Fig. 3.12 shows the total amount of power consumption of the four different schemes over five runs of simulation. Although the proposed scheme spens the least amount of power during operation, the difference among the four schemes is not significant. However, considering the unmet demand shown in Fig. 3.10, the energy efficiency differs to a large degree between different schemes as shown in the following Fig. 3.13.

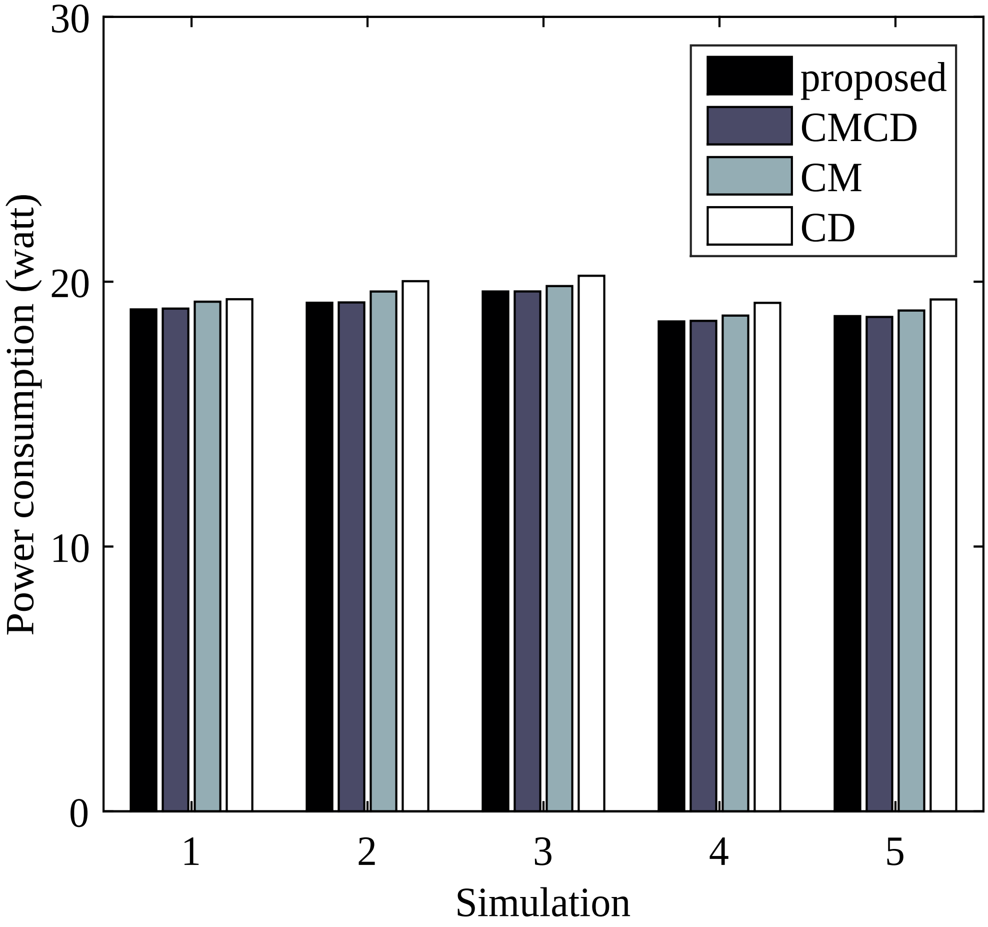

Figure 3.12: Total power consumption with respect to different ways of handling uncertainties over five runs of simulation.

Instead of considering only the power consumption, Fig. 3.13 takes both profit and power consumption into account and depicts the amount of power consumed to make a unit profit (i.e., a dollar) as an energy-efficiency metric. Clearly, CMCD consumes much more power compared to the rest three methods to make a unit profit, resulting in a low energy efficiency. Given that CMCD has suffered most from the unmet demand, it has made the least profit from providing service to users, while paying a significant penalty for unmet demand. On the other hand, CM, CD and the proposed methods have yielded a relatively small amount of unmet demand while consuming a similar amount of power to CMCD, resulting in a much higher profit and the energy efficiency

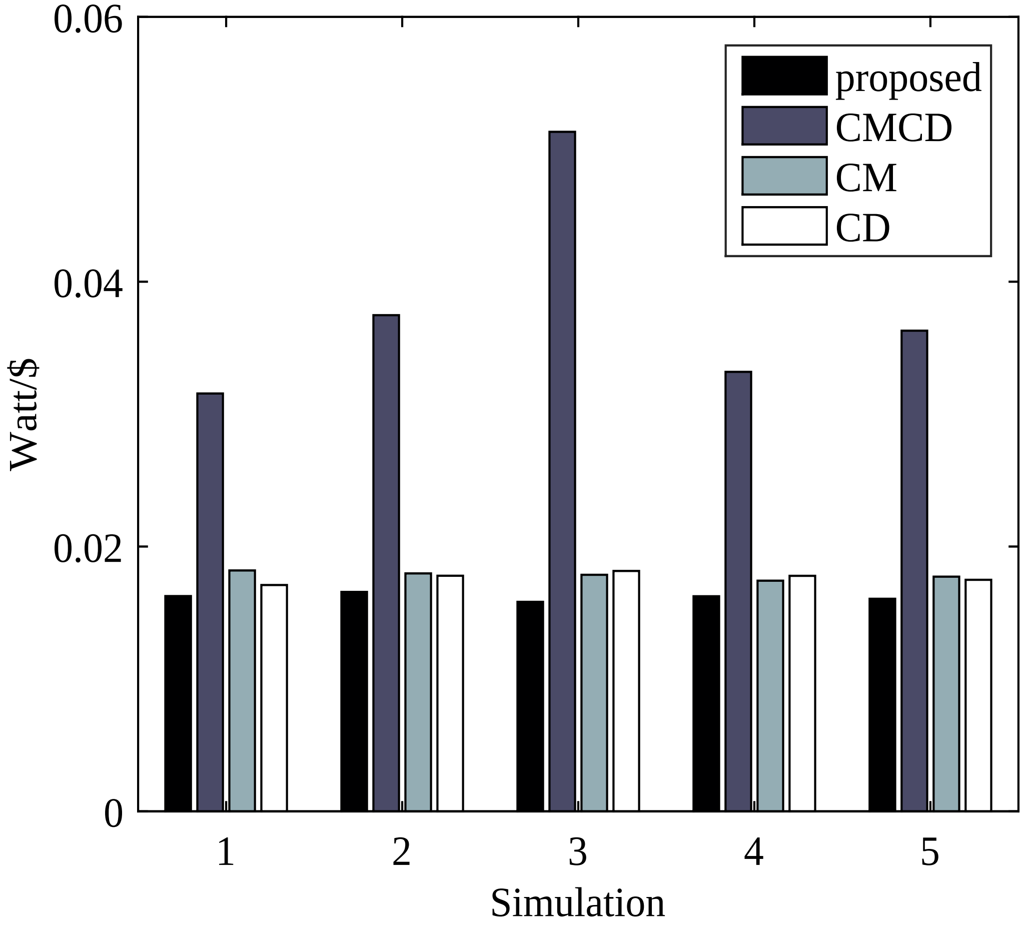

Figure 3.13: Comparison of the power consumed to make a unit profit ($) with respect to different ways of handling uncertainties over five runs of simulation.

than CMCD. Still, the proposed scheme outperforms the rest since it has recorded the least among the four methods in both power comsumption and unmet demand.

In what follows, we compare the per-stage profit of the four schemes, and study how the decisions made by each method affects the revenue. Fig. 3.15 shows the `Stage-1` profit of the four schemes. The profit in this stage cannot be positive since MNO_1 does not provide any service to users. Instead, it pays the expense for both operating its networking resource (i.e., BBU pool) and borrowing resource from MNO_2. The expense for operation is not much different across the four scheme, and thus, the different in the `Stage-1` profit mainly comes from the amount of resource borrowed from MNO_2. The proposed scheme, CM and CD are fully or partially aware of the uncertainties in the network and thus borrow a certain amount of resource from MNO_2. However, CMCD is not, and thus, it concludes that the users' demand can be satisfied solely by using MNO_1's networking resource. As a result, CMCD spends the least amount of expense in `Stage-1`.

The effects of the decision made in `Stage-1` does not become evident in `Stage-2` yet. The Fig. 3.16 shows the `Stage-2` profit, showing that the difference in the profit made in `Stage-2` is trivial among the four schemes. In fact, CMCD achieves the most profit, implying that CMCD is actively and heavily using the MNO_1's networking resource to fulfill the service demand from users. It would be an optimal decision if there was no uncertainty on the network, but it is not the case as shown in the following Fig. 3.17.

Finally, the Fig. 3.17 shows the `Stage-3` profit, on which stage the traffic offloading to MNO_2 occurs and the penalty to the unmet demand is measured. Since there is no scheme that completely fulfills the service demand, all the four scheme need to pay the price for the unmet demand which affects the overall revenue. However, only CMCD has yielded a negative profit in `Stage-3` since it did not reserve any resource at MNO_2 while having much unmet demand. Thus, all the service demand that has not satisfied at the end of `Stage-2` will be considered as unmet demand, and for which MNO_1 has to pay the penalty. However, the rest three have resulted in nonnegative profit in `Stage-3`. That is, although each of the three schemes has a strongly positive amount of penalty to be paid for the unmet demand, it is offset by the profit from providing service to users through the MNO_2's network infrastructure. In addition, the proposed method has successfully made a strongly positive profit in `Stage-3` over all runs of simulation, showing the advantage of being well aware of the network uncertainties. In sum, being well aware of the network uncertainties can yield a more profitable and energy-efficient solution compared to others that are completely or partially not.

3.5.4 Effect of Different Operation Rules

In this section, we study the effects of different *operation rules* on the performance. The proposed method allows a negligible amount of QoS reduction at the expense of penalty, which is done by introducing a nonnegative slack variable $\widetilde{r}_s^u$ as in Eq. 3.3g and Eq. 3.3h. Also, by allowing traffic offloading to other MNOs, the proposed method minimizes the amount of unmet demand. In order to study the effect of such operation rules, we compare the proposed method to other models with different operation rules. Please note that the methods to be used for comparison in this section are aware of the network uncertainties as much as the proposed method is. The models to be considered in this section are:

- **PSD** (perfect service demand) which has to perfectly satisfy service demand.
- **NoOFL** (no offloading) which does not allow any traffic offloading to other MNOs.
- **proposed** which allows both service outage (with penalty) and traffic offloading.

We evaluated the three different rules with respect to the same evaluation criterion as in Section. 3.5.3. We have carried out five runs of simulation, and the averaged results are summarized in Table 3.3.

As it can bee seen in the table, PSD resulted in an infeasible solution over all runs of simulation, and thus it could not provide any service to users. Due to the mobility and varying service demand of users, MNO_1 cannot always fully satisfy the users' demand. This implies that MNOs cannot put a PSD-like rule into effect since it might cause a service outage to the entire system when there is no feasible solution.

Also, Table 3.3 shows that the proposed method much outperforms NoOFL. The achieved profit of the proposed method is almost twice that of NoOFL. The main reason for such a significant difference in profit is the penalty to pay for unmet demand in `Stage-3`. In contrast to the proposed method which made a positive profit in `Stage-3`, NoOFL resulted in a large expense in the same stage mainly due to the large amount of unmet demand. Considering that both methods used similar amount of power during operation, the proposed method significantly outperforms NoOFL in terms of the power consumption per watt. Since NoOFL does not allow any offloading to MNO$_2$, its optimal solution and the resulted performance are similar to that of CMCD. It is noteworthy that CMCD and NoOFL are totally different in terms of the awareness of the uncertainties. However, due to the fact that both do not or cannot leverage any resource from MNO$_2$, the performance of one becomes similar to that of the other. This result proves the significance of allowing traffic offloading which is as important as considering the uncertainties in the network.

In sum, a strict operation rule that compels an MNO to fully satisfy the users' service demand should not be considered in practice since it can cause the entire network to be unavailable for a long period of time. Also, being capable of traffic offloading significantly increases the system performance, since a network can make use of more networking resource than it actually owns. In other words, sharing the network resource among different MNOs is an effective way of enhancing the network performance.

3.5.5 Time Complexity: Runtime

For evaluation, we have used a laptop with Intel$^{®}$ Core^{TM} 2 Duo P8700 2.53 GHz CPU and 4 GB memory. As illustrated in Algo. 2, the resource allocation algorithm iterates for $|\mathcal{M}_1| - 1$ times, and within each iteration it solves DEP (i.e., P. 3.4) for a reduced scenario tree $\mathcal{T}'$. In the case of CMCD (or the expected value problem, referred to as EV), it took only 4.56 seconds (standard deviation is 0.83) to find an optimal solution. Since EV assumes the mean values for all uncertainty parameters, the problem size as well as the dimension of the decision variables is small and thus, it can be solved instantly. On the other hand, the stochastic program (i.e., DEP) took 87.64 seconds to solve (standard deviation is 2.34), which is much larger than that for EV. The difference in cputime comes from the difference in problem size and the dimension of the decision variables. To be specific, DEP takes much more data sets into account and thus, has much larger set of decision variables. As a result, the stochastic approach results in a better solution at the expense of cputime. In our implementation, the complexity of other methods that are not mentioned here is identical to that of the stochastic problem, and thus the cputime results of those are omitted.

3.5.6 Stability and Quality of Solution

As aforementioned, we have used a sampling method as a part of the solution procedure to reduce the problem size and thus a computer solver can solve the optimization problem. To be specific, out of all possible combinations of the uncertainty parameters describing the connectivity and service demand information, we draw only a subset at uniform random to construct a reduced scenario tree $\mathcal{T}'$. Thus, the optimal solutions from different scenario trees can vary depending on the construction of the corresponding scenario trees. In this regard, it is important to evaluate the stability and quality of the solution. In this section, we first show that the solutions from different scenario trees have in-sample and (weaker) out-of-sample stability, meaning that the solution does not depend much on the construction of scenario trees. Also, by evaluating the solution quality, we show that the upper bound of the error caused by the sample average approximation method is small. Interested readers can refer to [77] for an in-depth discussion on the stability and quality of solution for stochastic programming.

For notational simplicity, let $\phi(\mathbf{x}; \mathcal{T}_i)$ be the optimization problem, where ϕ is the DEP instance, $\mathbf{x}$ is the set of decision variables, and $\mathcal{T}_i$ is a scenario tree indexed by i. By solving the problem, we get the optimal solution $\hat{\mathbf{x}}_i = \arg\max_{\mathbf{x}} \phi(\mathbf{x}; \mathcal{T}_i)$. Also, we have an objective value $\rho_{ij} = \phi(\hat{\mathbf{x}}_i; \mathcal{T}_j)$ when the DEP instance is evaluated with $\hat{\mathbf{x}}_i$ for $\mathcal{T}_i$. In this section, we have used four different scenario trees to evaluate the solution stability and quality.

3.5.6.1 In-Sample Stability

The solution procedure has an in-sample stability if the following is satisfied [77]: $\phi(\hat{\mathbf{x}}_i; \mathcal{T}_i) \approx \phi(\hat{\mathbf{x}}_j; \mathcal{T}_j)$, which happens when the selection of scenario tree does not affect the optimal objective value much. Thus, if we have an in-sample stability, we can choose any scenario tree when solving a stochastic program, and the corresponding optimal solution will be comparable to the ones from other scenario trees. We have used the Jain's fairness index [84], $\mathcal{J}(\rho_1, \rho_2, \cdots, \rho_n) = \frac{(\sum_{i=1}^{n} \rho_i)^2}{n \times \sum_{i=1}^{n} \rho_i^2}$, to check how much are the objective values from different scenario trees close to each other. Out of four scenario trees we have constructed, we have a fairness index of $\mathcal{J}(\{\rho_{ii} | \forall i = 1, 2, 3, 4\}) = 0.9999$

which is close to 1. That is, the optimal objective values from five different scenario trees are close to each other. It implies that no matter how each scenario tree is configured, the optimal value (i.e., optimal profit) from such a scenario tree will not much differ from the others from different scenario trees.

3.5.6.2 Weaker Out-of-Sample Stability

The weaker out-of-sample stability checks if the solution from a scenario tree is still a good solution to other scenario trees as well. We have a weaker out-of-sample stability, if the following holds [77]: $\phi(\hat{\mathbf{x}}_i; \mathcal{T}_j) \approx \phi(\hat{\mathbf{x}}_j; \mathcal{T}_i)$. The Tab. 3.5 show the weaker out-of-sample stability test results. The $(i, j)^{\text{th}}$ entry in the table is the results of $|\rho_{ij} - \rho_{ji}|$. The mean of the values above the diagonal is 30.6819 which is merely 2.64% of the mean of ρ_{ij} for all i, j where $i \neq j$. Therefore, the solution procedure has a weaker out-of-sample stability.

3.5.6.3 Solution Quality

Since the solution procedure has both in- and out-of-sample stability, we then study the solution quality. By definition [77], the quality of a given solution $\hat{\mathbf{x}}^i$ corresponds to the optimality gap which is defined as $\text{err}(\hat{\mathbf{x}}_i) = \max_{\mathbf{x}} \phi(\mathbf{x}; \xi) - \phi(\hat{\mathbf{x}}_i; \xi)$. However, it is impossible to find the gap because of the vast size of the original problem. Instead, we calculate a statistical estimate of the error of a particular solution $\hat{\mathbf{x}}_i$, which is: $\text{err}(\hat{\mathbf{x}}_i) \lesssim \frac{1}{n} \sum_{j=1}^{n} [\max_{\mathbf{x}} \phi(\mathbf{x}; \mathcal{T}_j) - \phi(\hat{\mathbf{x}}_i; \mathcal{T}_j)]$, where $n = 4$ and $\leq$ can be used instead of $\lesssim$ as $n \to \infty$. The Tab. 3.5 shows the stochastic upper bound on the error for each solution. Compared to the objective values which are greater than or equal to 1114.3297 in any cases, the error is small, implying a high quality of the solutions.

3.6 Conclusion

In this chapter, we have studied optimal resource allocation in H-CRAN under uncertainty. In order to maximize profit as well as to minimize the power consumption, the proposed method optimizes the set of resources constituting H-CRANs, such as BBU pool, channels and BSs given fronthaul and QoS requirements. Also, by allowing different network operators to share their resources by means of traffic offloading/roaming, the proposed method has further increased the profit. By taking a multi-stage SP approach, the uncertainties in both users' mobility and their service demand are taken into account and handled for resource optimization. The evaluation results show that the proposed method can increase the profit, energy-efficiency and QoS compared to the methods that do not or partially consider the uncertainties. In addition, we have shown that the traffic offloading among different MNOs can further increase the cost-efficiency of H-CRANs.

Table 3.1: Summary of notations used in Chapter 3

	Sets
$\mathcal{I}_1$	Index set of RRHs, $\{1, 2, \cdots, i_1, \cdots\}$
$\mathcal{I}_2$	Index set of AP, $\{1, 2, \cdots, i_2, \cdots\}$
$\mathcal{M}_1$	Index set of channels MNO$_1$ can access to, $\{1, 2, \cdots, m_1, \cdots\}$
$\mathcal{M}_2$	Index set of channels MNO$_2$ can access to, $\{1, 2, \cdots, m_2, \cdots\}$
$\mathcal{S}$	Index set of scenarios, $\{1, 2, \cdots, s, \cdots\}$, $\lvert\mathcal{S}\rvert = \lvert\mathcal{T}\rvert$
$\mathcal{T}$	Complete scenario tree (i.e., complete set of scenarios)
$\mathcal{U}$	Index set of UEs, $\{1, 2, \cdots, u, \cdots\}$
	Uncertainty parameters
$\mathbf{r}$	Users' service demand
$\mathbf{V}$	Accessibility indicator between UEs and RRHs
$\mathbf{W}$	Accessibility indicator between UEs and APs
ξ	Description of accessibility and demand, $(\mathbf{V}, \mathbf{W}, \mathbf{r})$
	Deterministic parameters
α	Conversion parameter from serviced bits to revenue
β	Offloading cost
c_{i_1}	Backhaul constraint on RRH i_1
d_{i_2}	Association capacity of AP i_2
δ	OPEX of BBU pool (i.e., computing resource)
ϵ	Computing resource reservation cost
$f^u_{i_1,m_1}$	Bandwidth of channel m_1 between UE u and RRH i_1
$g^u_{i_2,m_2}$	Bandwidth of channel m_2 between AP i_2 and UE u
γ	Penalty for unmet demand
K_1	MNO$_1$'s conversion parameter for processing capacity
K_2	MNO$_2$'s conversion parameter for processing capacity
o_{i_1}	Association capacity of RRH i_1
τ^M	Per-channel transmission power of MBS
τ^R	Per-channel transmission power of RRH
θ	Energy cost
z_{i_2}	Aggregate bandwidth capacity of AP i_2
	Decision variables
x_{i_1}	Fraction of BBU pool allocated to RRH i_1, where $x_{i_1} \in \mathbf{x}$
y	Fraction of computing resource reserved at MNO$_2$'s BBU pool
$\boldsymbol{\pi}$ or $\overline{\boldsymbol{\pi}}$	MNO$_1$'s channel allocation vector for MBSs or RRHs
$\mathbf{A}_s$	UE, BS (MBS or RRH) and channel mapping in scenario s
$\mathbf{B}_s$	UE, AP and channel mapping in scenario s
$\widetilde{r^u_s}$	Unmet demand of UE u in scenario s, where $\widetilde{r^u_s} \in \widetilde{\mathbf{r}}_s$

Table 3.2: Network parameters

$\lvert\mathcal{I}_1\rvert$	4	ϵ	2δ
$\lvert\mathcal{I}_2\rvert$	4	δ	\$2.608/hour [82]
$\lvert\mathcal{M}_1\rvert$	5	K_1	300Mbps
$\lvert\mathcal{M}_2\rvert$	5	K_2	300Mbps
$\lvert\mathcal{U}\rvert$	15	α	\$20/300MB/mo [83]
$\lvert\mathcal{T}'\rvert$	30	θ	\$0.0897 per watt-hour [72]
o_{i_1}	3 ($\forall i_1 \in \mathcal{I}_1$)	γ	3α
d_{i_2}	3 ($\forall i_2 \in \mathcal{I}_2$)	β	\$15/500MB [56]
z_{i_2}	30Mbps ($\forall i_1 \in \mathcal{I}_1$)	c_{i_1}	30Mbps ($\forall i_2 \in \mathcal{I}_2$)

Table 3.3: Comparison of the mean performance out of five runs of simulation.

Criterion	Proposed	PSD	NoOFL
Achieved profit (\$)	1173.7990	n/a	522.5996
Mean unmet demand (Mbps)	0.1000	n/a	1.5735
Total power consumption (Watt)	18.9981	n/a	18.9949
Power use per profit (Watt/\$)	0.0162	n/a	0.0372
Fraction of resource reservation (y)	0.7241	n/a	0.0
Stage-1 Profit (\$)	-1.3259	n/a	-0.2775
Stage-2 Profit (\$)	1104.0434	n/a	1115.8850
Stage-3 Profit (\$)	71.0812	n/a	-593.0078

Table 3.4: Results for the weaker out-of-sample stability.

$\lvert\rho_{ij} - \rho_{ji}\rvert$	$j = 1$	$j = 2$	$j = 3$	$j = 4$
$i = 1$	0	42.7756	50.6854	19.4565
$i = 2$	42.7756	0	19.5922	20.0975
$i = 3$	50.6854	19.5922	0	31.4837
$i = 4$	19.4565	20.0975	31.4837	0

Table 3.5: Stochastic upper bound on the error for each solution.

Solution	$\hat{\mathbf{x}}_1$	$\hat{\mathbf{x}}_2$	$\hat{\mathbf{x}}_3$	$\hat{\mathbf{x}}_4$
Error	28.9169	8.4497	8.4824	13.2394

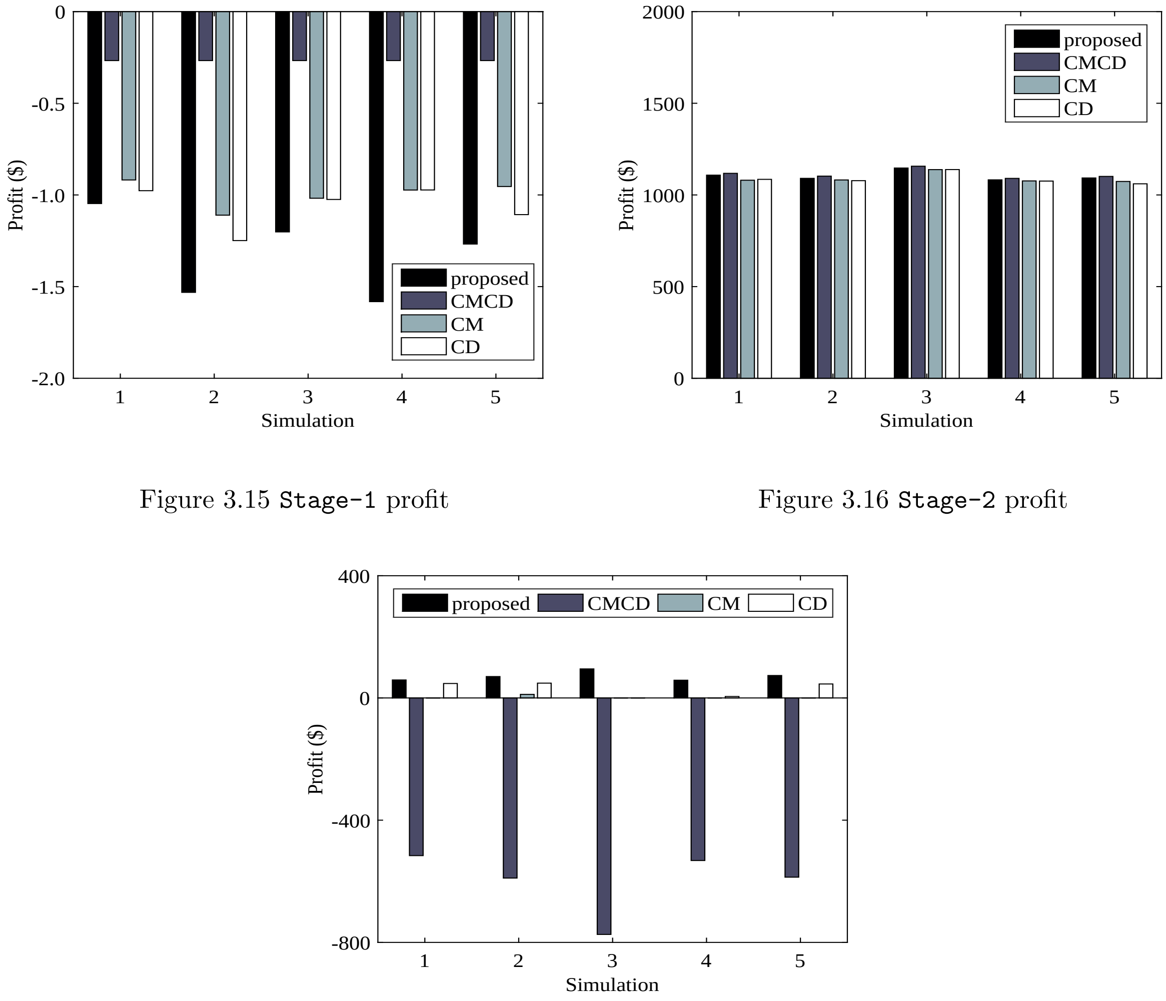

Figure 3.15 `Stage-1` profit

Figure 3.16 `Stage-2` profit

Figure 3.17 `Stage-3` profit

Figure 3.18: Comparison of the per-stage profit with respect to different ways of handling uncertainties over five runs of simulation.

CHAPTER 4. ENHANCED POWER SAVING MECHANISM FOR NEXT GENERATION WLANS: LARGE-SCALE 802.11AH INTERNET OF THINGS NETWORKS

4.1 Summary

We consider the power saving mechanism for 802.11ah on large-scale sensor networks. Power saving mode allows sensor nodes to switch to low-power state while APs buffer incoming frames for them. Also, 802.11ah introduced TIM and page segmentation scheme to cope with a large number of nodes. Although it is a powerful tool for reducing contention and increasing sleep intervals, the advantage comes at the expense of additional energy waste, which becomes exacerbated as the number of nodes increases. In this chapter we aim at minimizing such energy waste and enhancing energy efficiency which is critical to large-scale 802.11ah networks. We propose a method that selectively and dynamically changes the membership of nodes and rearranges their traffic to maximize overall sleeping intervals without causing delay to data delivery. To this end, we propose a temporary membership change scheme and a traffic scheduling algorithm which reduce overall power consumption. Also, to make the problem tractable and scalable, we apply a relaxation technique and devise a low-complex scheduling algorithm, respectively, of which performance is comparable to optimum. Evaluation results show that the proposed scheme can enhance energy efficiency by decreasing the number of active nodes by up to 37.8% compared to 802.11ah. The complete version of this chapter has been published in [85].

4.2 Introduction

A long with the continuous development of the communication devices and networking technology, the idea which is to connect physical wireless sensor nodes located at the surrounding area and thus, to provide a diverse range of applications has received much attention. Among various tech-

Table 4.1: Abbreviations and acronyms used in Chapter 2

AID	Association Identifier
AP	Access Point
DL/UL	Downlink/uplink
(D)TIM	(Delivery) Traffic Indication Map
IoT	Internet of Things
MAC	Medium Access Control
RAW	Restricted Access Window
STA	Station, i.e., (sensor) node/device
TIM STA	a STA that shows its paged status in TIM

niques in connecting a large number of devices to the Internet, IoT[1] has recently drawn attention from academia and industry, although it was first introduced back in 1998 [89]. IoT is expected to play a remarkable part in many applications, such as environmental monitoring, smart building, health care, home automation, disaster alerting and ambient assisted living [89] [90] [93], to name a few. Also, its potential for practical use has been shown in many literatures, including IoT testbeds [87] and the off-the-shelf market products [88].

One of the major technical challenges in realizing IoT applications is an absence of unified communication technology for sensor nodes. Therefore, different applications have been using different networking systems to connect devices (e.g., Zigbee, 802.15.4, 6LoWPAN, Bluetooth, cellular networks, etc. [91]) for their own good or specific missions [86]. To this end, IEEE 802.11ah Task Group was formed in 2010 aiming at providing a unified solution for connecting and integrating a large number of (possibly heterogenous) battery-powered devices at the license-exempt band below 1 GHz, excluding the TV White Space bands.

Although a part of the specification [9] is inherited from the previous IEEE 802.11 standards, e.g., 11n and 11ac, a long transmission range (∼1km), supporting a large number of STAs (up to ∼8,000), coexistence/integration with 802.15.4 devices, and an enhanced power saving are the ones that uniquely distinguish 802.11ah from the rest of the IEEE 802.11 family; interested readers can refer to [8] and [92] for system requirements and use cases, respectively. In other words, unlike the

[1] Abbreviations and acronyms that frequently appear in this chapter are summarized in Table 4.1.

previous 802.11 standards that are designed to provide high data rates for small-/medium-sized networks, 802.11ah is intended to handle a large number of battery-limited STAs per AP with a low data rate [93] and a large coverage for an extended lifetime, which is why energy efficiency has to be one of the primary concerns.

As expected, 802.11ah inherited the power saving mechanism from its predecessors to prolong the lifetime of STAs. In addition, *TIM and page segmentation*, combined with a restricted access mechanism, is introduced to efficiently cope with a large number of STAs; to be specific, STAs are clustered into multiple groups, and each of which is assigned a dedicated time interval for channel access. By efficiently limiting the level of contention, STAs can spend less time on channel access and more in low-power state. For a STA to retrieve buffered data from AP, it periodically wakes up to receive DTIM, which is a group-wise paging status, and if paged, it wakes up on their assigned time interval to listen to TIM, which is a STA-wise paging status. In spite of its advantage, however, the paging-based power saving mechanism causes an unnecessary energy waste problem. A STA that does not have any buffered data has to wake up if the group to which the STA belongs is paged, and the energy waste from such unnecessary wake-ups becomes exacerbated as the number of STAs increases.

Previous studies concerning power consumption [104] [106] [109] have focused on the energy efficiency mainly in regard to the channel access. This approach can be beneficial if STAs are backlogged and actively trying to access the channel for most of the time, but that is not likely to be the case, in general, for 802.11ah sensor networks [91] [92] [93]. Rather, maximizing both the number of sleeping STAs and their sleep intervals needs to be a top priority to enhance energy efficiency for a large-scale sensor network with low traffic rate. It has not been noticed that, as the number of STAs increases, 802.11ah power saving and paging scheme might cause a frequent, unnecessary energy waste for the aforementioned reason, which is the motivation of this chapter. In this regard, our goal is to identify and address the unnecessary energy waste problem of 802.11ah power saving mechanism. To this end, we propose a novel solution that maximizes the overall energy efficiency while being compliant with 802.11ah.

In this chapter, we propose a dynamic membership change mechanism for a large-scale 802.11ah wireless sensor network to prolong the lifetime of STAs by increasing their sleeping periods. Specifically, the proposed mechanism reduces the number of STAs that have to unnecessarily wake up under the 802.11ah power saving mechanism. To this end, we formulate an optimization problem that establishes an additional membership relation. Its solution provides an optimal strategy that allows some selected STAs to temporarily change their membership such that the following scheduling algorithm can maximize the number of sleeping STAs. Also, we apply a relaxation technique to the optimization problem to make it tractable, while keeping the optimality gap small. What follows is to propose a low-complexity traffic scheduling algorithm that maximizes the sleeping intervals for STAs. Its performance in terms of the number of STAs in low-power mode is comparable to the optimal scheduling scheme, while incurring much less overhead. The proposed idea does not depend on or assume any particular grouping of STAs. Also, it makes use of the same TIM/AID structure as defined in 802.11ah [9] and does not require any significant changes. Thus, the proposed method is compatible with 802.11ah with little modification. Finally, the proposed traffic scheduling algorithm is lightweight, making it suitable for online processing.

The rest of the chapter is organized as follows. A brief introduction to IEEE 802.11ah MAC and the motivation of the research is given in Section 4.3. The following Section 4.4 introduces the network system model and assumptions. Section 4.5 and 4.6 presents the proposed *Secondary AID Assignment* and the *Traffic Scheduling* algorithm, respectively. The evaluation and comparison results are presented in Section 4.7, and finally, we conclude the chapter by Section 4.8.

4.3 Preliminaries

In this section, we first provide a brief introduction to 802.11ah MAC. In what follows, we present an illustrative example scenario that motivated this research.

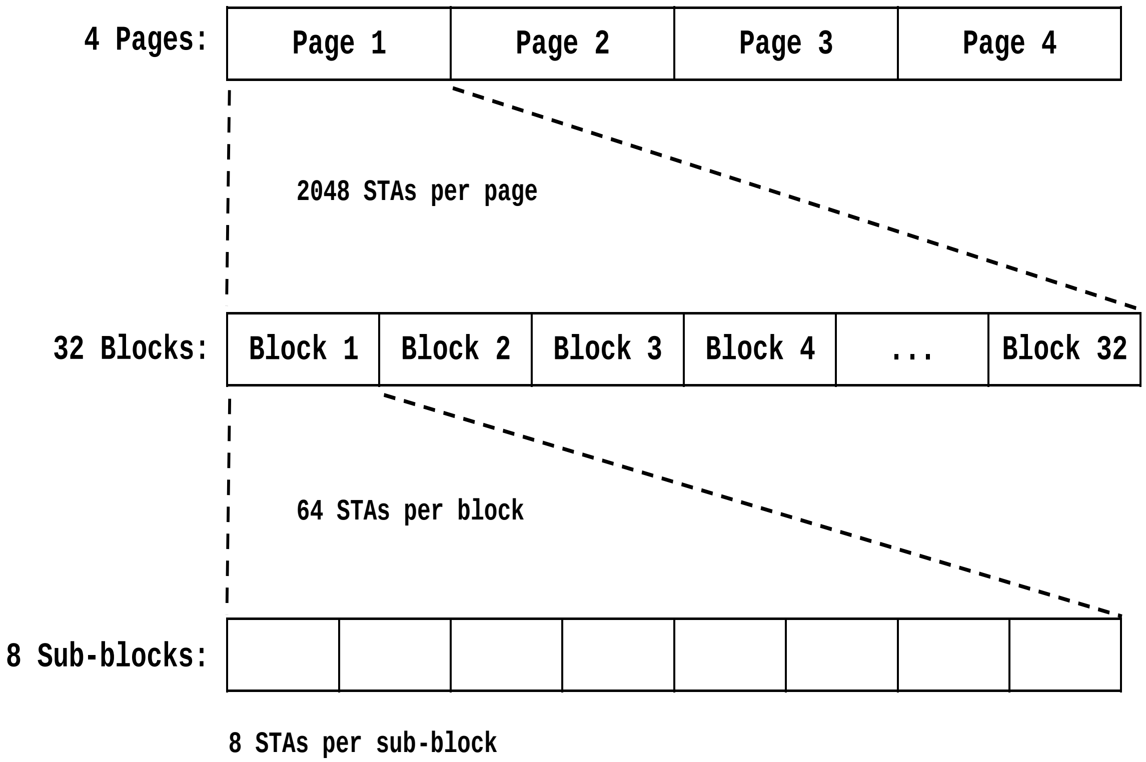

Figure 4.1: Illustration of the 802.11ah TIM structure.

4.3.1 802.11ah MAC: TIM and Page Segmentation

In this study, we focus on the 802.11ah TIM STAs whose *paged* status is included in TIM; interested readers can refer to [9] [91] for the detailed description of the 802.11ah MAC features. The term *paged* in this chapter refers to the case that there is at least one data buffered at AP. A STA is paged (by TIM) if an AP has pending traffic for the STA, and in this case, the TIM group to which the STA belongs has to be paged (by DTIM) as well.

TIM and page segmentation have been introduced in 802.11ah to effectively manage a large number of STAs, reduce the level of contention, and enhance the energy efficiency in a structured manner. By limiting the number of STAs that are allowed to participate in contention at the same time, STAs spend less time in channel access and more in low-power state. TIM is structured in a 3-level hierarchy as shown in Fig. 4.1. Throughout the chapter, we assume the same parameters for the number of pages and blocks as in [9]. As shown in Fig. 4.2, the structure of AID is

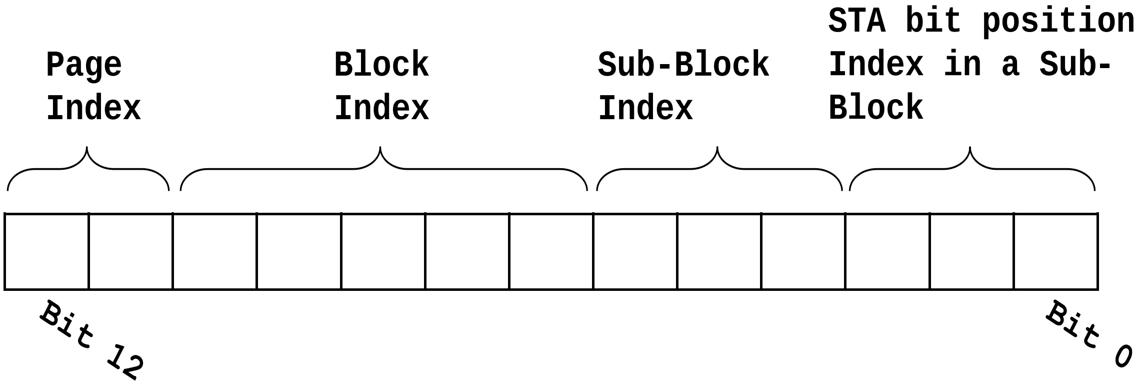

Figure 4.2: Illustration of the 802.11ah AID structure.

closely related to that of TIM by having a 3-level hierarchical structure, i.e., page/block/sub-block. A set of STAs whose AIDs are within a certain range belong to the same TIM block (referred to as *group* in this chapter), and they are assigned to the same time period during which they are exclusively allowed to access the channel. In addition to the structured TIM/AID as well as the grouping of STAs, both power saving mechanism and page segmentation enable an efficient power and transmission management.

The power saving mode allows STAs to turn off their communication interfaces for a certain period of time to reduce energy consumption and, in the meantime, an AP buffers incoming traffic for them. TIM stations periodically wake up to receive DTIM/TIM, which lets TIM STAs sleep as much as possible and wake up only when they are allowed to contend for the channel. STAs in the same TIM group wake up during their designated time period, and thus their contention is not interfered by the rest TIM groups. Therefore, the level of contention for accessing the channel becomes limited, while the sleeping interval increases. Time is slotted into multiple DTIM intervals, each of which is further divided into multiple TIM intervals. At the beginning of each DTIM interval, all TIM STAs must wake up and listen to DTIM which is a group-wise page status map. However, it does not specify which STA(s) is paged. Therefore, every STA in the paged TIM group needs to wake up on its TIM interval (i.e., the time interval for the group to which the device belongs) to see if it actually has any buffered traffic at the AP; on the other hand, STAs in the non-paged TIM groups remain asleep.

4.3.2 Motivation

Although the TIM-based page segmentation scheme has many advantages, it might incur an undesired energy waste for what is called *unnecessary wake-up* in this chapter. As aforementioned, STAs in low-power state must wake up to receive DTIM and TIM, if their group is paged. For those STAs that expect frequent traffic arrivals, waking up on their TIM interval is likely to be *necessary*; in other words, it is likely that the AP has buffered data for them. On the other hand, a STA which rarely or occasionally receives traffic will see its wake-up as *unnecessary* in most cases; in other words, it wakes up only because one or more other STAs in the same TIM group is paged, and thus its wake-up will be a waste of energy (i.e., there is no buffered traffic for it). To make matters worse, the unnecessary wake-up problem becomes exacerbated as the number of STAs increases because the probability of a group being paged increases with the aggregate traffic rate of the group. As a result, the energy efficiency of each individual STA as well as that of the entire network is degraded.

Although a few number of such unnecessary wake-up events might consume a negligible amount of power, considering that the expected lifetime of an 802.11ah STA is long (i.e., from months to years [91]) and the number of STAs on the network is large (up to ∼8,000), a STA which is frequently exposed to such a case will suffer from a severe energy waste in the long run. For example, let the duration of a single TIM group be 200ms [91] and the number of TIM groups in total is 32, which results in 6.4s of one DTIM interval. Assuming a year of expected lifetime, each STA will encounter approximately 5×10^6 times of DTIM intervals. On each interval, a STA that unnecessarily wakes up will waste σ amount of more energy compared to staying in low-power state. The additional energy waste σ amounts to: $2 \cdot P_{ST} + P_{ACTIVE}(t) + P_{RCV}(t) - P_{SLEEP}(t)$, where P_{ST}, P_{ACTIVE}, P_{RCV} and P_{SLEEP} are the power consumption for making state transition, staying in active state, signal reception and decoding, and staying in sleep state, respectively, and t is the amount of time spent. For each individual STA, the increase in the probability, p, of making unnecessary wake-ups results in wasting approximately $5 \times 10^6 p\sigma$ amount of energy, which is mainly caused by having more STAs in the same group. Obviously, the total amount of energy waste from the entire network significantly increases for the same reason. In order to reduce such unnecessary wake-ups and thus, to increase the energy efficiency, we propose a scheme that enables both a temporary membership change and traffic scheduling. The following example shows how the proposed idea can efficiently achieve the goal without requiring much of the changes to the 802.11ah operation.

Let us assume a simple TIM and AID structure where there is only one page and two TIM groups, and each TIM group can associate with up to 4 STAs. On the network, seven STAs, from n_1 to n_7, are associated and grouped as: $TIM_1 = \{n_1, n_2, n_3\}$ and $TIM_2 = \{n_4, n_5, n_6, n_7\}$. Here, TIM_2 group is fully occupied, while TIM_1 has one unused (AID) slot. Right before the beginning of the i-th DTIM interval, the AP has two buffered data, one for $n_1 \in TIM_1$ and the other for $n_7 \in TIM_2$. Since both TIM groups have at least one pending data for each at the AP, the paged status or block bits (i.e., indication of the presence of the buffered traffic for each TIM group) for both TIM groups need to be set in DTIM. At the beginning of the TIM period for each group, all STAs in the corresponding group must wake up and receive TIM since their group is paged in DTIM. In this simple example, five stations experience unnecessary wake-up events.

On the other hand, let us consider the case where the AP additionally assigns the empty slot in TIM_1 to n_7 so that n_7 can be regarded as a member of TIM_1 as well. For the same buffered traffic, the AP would set the block bit for TIM_1 only, since all the pending data at the AP can be delivered by just waking up TIM_1. In this setting, the number of unnecessary wake-ups is only 2, n_2 and n_3, and the remaining STAs (i.e., all STAs in TIM_2 except n_7) can stay in low-power state. It is worth mentioning that as it can be seen in Section. 4.7, the gain of the proposed method increases with the network size; in other words, the unnecessary wake-up problem deteriorates as the number of STAs on the network increases. This temporary membership change approach does not require a STA to re-associate or any change in the management/organization of the entire STAs. In addition, as discussed in Section. 4.6, it does not increase the expected delay to data delivery.

4.4 Network Model and Assumptions

We consider a single-hop 802.11ah sensor network where all STAs communicate directly with an AP on a shared wireless channel. We assume that traffic arrivals to STAs (e.g., actuation or control messages in DL) follow the Poisson distribution. The rate of each arrival is assumed to be known or can be learned by either an explicit notification or monitoring the traffic, respectively. Given the traffic rates, we classify the STAs into two groups; one is called *sensory* whose traffic arrival rates are small, and the other is called *controllable* whose incoming traffic rates are relatively large. Sensory STAs are the ones whose main task is sensing the environment or metering the usage of resources, and then they send the information in the UL. Therefore, their incoming traffic rates are low. On the other hand, the controllable STAs are the ones that frequently receive either control or actuation messages in the DL so that they can perform some designated actions/tasks or missions. We assume that the majority of STAs are sensory. However, we do not assume anything about the number of devices on the network and how the association/grouping has been made.

In the 802.11ah network, any STAs that want to join the network should make an association with an AP. As a result, STAs are given AIDs which also indicate to which TIM group each STA belongs. Let us denote this association as *primary association*, and the corresponding AID as *primary AID* (or P-AID). After the primary association is made, the proposed method runs the following two programs: *Secondary AID Assignment* and *Traffic Scheduling*. The Secondary AID Assignment program utilizes the unused AIDs, and makes additional associations between *some* STAs and groups so that those STAs can also associate with other groups in addition to the group they initially belong to by primary association. In other words, the Secondary AID Assignment program establishes a secondary membership, called *secondary association*, so that some STAs have two AIDs and thus, they can switch between the groups. The second AID to be given to some STAs as a result of the secondary association is called *secondary AID* (or S-AID for short). Once the secondary AID assignment is made, the following traffic scheduling program checks if there is a chance to reduce the number of STAs that have to wake up unnecessarily. As illustrated by example in Section. 4.3, if the pending traffic at the AP can be completely delivered by waking up

Table 4.2: Summary of notations used in Chapter 4

M	Number of TIM groups
$\mathcal{G}$	Set of TIM groups, $\{g^{(1)}, g^{(2)}, ..., g^{(M)}\}$
$n_s^{(i)}$	Number of sensory STAs in $g^{(i)}$
$\mathcal{S}^{(i)}$	Set of sensory STAs in $g^{(i)}$, $\{s_1^{(i)}, s_2^{(i)}, ..., s_{n_s^{(i)}}^{(i)}\}$
$n_c^{(i)}$	Number of controllable STAs in $g^{(i)}$
$\mathcal{C}^{(i)}$	Set of controllable STAs in $g^{(i)}$, $\{c_1^{(i)}, c_2^{(i)}, ..., c_{n_c^{(i)}}^{(i)}\}$
$y^{(i)}$	Number of unused AIDs in $g^{(i)}$
$\Lambda^{(i)}$	Arrival rates of sensory STAs in $g^{(i)}$, $\{\lambda_1^{(i)}, \lambda_2^{(i)}, ..., \lambda_{n_s^{(i)}}^{(i)}\}$
$\mathcal{M}^{(i)}$	Arrival rates of controllable STAs in $g^{(i)}$, $\{\mu_1^{(i)}, \mu_2^{(i)}, ..., \mu_{n_c^{(i)}}^{(i)}\}$
η	Amount of energy waste from an unnecessary wake-up

less number of TIM STAs, then the AP re-arranges the traffic indication bits in both DTIM and TIM message by taking advantage of the S-AID.

4.5 Problem Formulation: Secondary AID Assignment

The first part of the proposed scheme is to establish secondary associations for *some* STAs if there are unused AIDs. Since the number of unused AIDs is limited, we cannot allow all STAs to establish the secondary association. In this regard, we propose to chose controllable devices and let them be assigned S-AIDs, since they are the ones that mainly cause frequent unnecessary wake-ups to other STAs. The classification of STAs into either sensory or controllable group depends only on their traffic arrival rates, which are, as aforementioned, known in advance or can be learned during operation. Before we proceed, please note that the notations used in the remainder of the chapter are summarized in Table 4.2.

4.5.1 Classifying Stations

The purpose of the classification is to make a list of controllable STAs which have higher traffic arrival rates, and eventually to pass the list to the Secondary AID Assignment program which

Algorithm 3 Classification Algorithm

```
 1: for each $g^{(i)}$, i=1 to M do
 2:    $R = \{r_1, r_2, \cdots, r_j, \cdots, r_{N^{(i)}}\}$                    // traffic rates
 3:    $R_{min} = \texttt{min}(R)$, $R_{max} = \texttt{max}(R)$
 4:    $S \leftarrow$ number of steps
 5:    $\Delta = (R_{max} - R_{min})/S$
 6:    $sum^*,\ thr^* \leftarrow \infty$
 7:    for t=1 to S-1 do
 8:       $thr = R_{min} + t \cdot \Delta$
 9:       $sum = \sum_{j=1}^{N^{(i)}} |thr - r_j|$
10:       if $sum < sum^*$ then
11:          $sum^* \leftarrow sum$, $thr^* \leftarrow thr$
12:       end if
13:    end for
14:    for each STA j=1 to $\texttt{N}^{(i)}$ do
15:       if $r_j \le thr^*$ then
16:          $\mathcal{S}^{(i)} = \mathcal{S}^{(i)} \cup \{\text{STA-}j\}$            // mark as sensory
17:          $\Lambda^{(i)} = \Lambda^{(i)} \cup \{r_j\}$
18:       else
19:          $\mathcal{C}^{(i)} = \mathcal{C}^{(i)} \cup \{\text{STA-}j\}$            // mark as controllable
20:          $\mathcal{M}^{(i)} = \mathcal{M}^{(i)} \cup \{r_j\}$
21:       end if
22:    end for
23:    $n_s^{(i)} = |\mathcal{S}^{(i)}|$, $n_c^{(i)} = |\mathcal{C}^{(i)}|$
24: end for
```

decides whether or not to assign an S-AID to each STA in the list. Although the classification procedure can be as complex as we want it to be, we have chosen a simple algorithm in order for it to be scalable. One extremely simple, yet working, solution is to mark all STAs as controllable. However, it maximizes the search space for the secondary AID assignment problem and thus, it rather increases the overall runtime. In this regard, we have designed a simple algorithm, Algo. 3, which finds a threshold point by which STAs are classified.

Algo. 3 iterates over each group and classifies STAs into one of the two groups, i.e., sensory or controllable. The set R (line 2) represents the traffic rates of all STAs in $g^{(i)}$, where $N^{(i)}$ is the number of STAs in $g^{(i)}$. The design parameter S (line 4) determines the number of iterations to search for the best threshold. The step size is equal to the range of traffic rates divided by S (line 5). On each iteration of the first inner for-loop (lines 7–13), the algorithm sets a threshold (line 8), calculates the sum of the distances between the threshold and all the traffic rates (line 9), and records the so-far the best threshold that minimizes the distance sum (lines 10–12). Then, in the later for-loop (lines 14–22), STAs are classified into either sensory or controllable by using the best threshold point.

4.5.2 Expected Power Consumption

As a reminder, the overall goal of the proposed method is to reduce the number of unnecessary wake-ups and thus, to reduce the overall energy consumption. In order to find a secondary association strategy that helps to reduce such wake-ups most, we first define the expected power consumption from the unnecessary wake-ups of sensory STAs[2]. Let us first define the following events: $A^{(i)}$:= there is at least one frame buffered at the AP for $g^{(i)}$ in time t (i.e., a DTIM interval); $a_j^{(i)}$:= there is at least one frame buffered at the AP for $s_j^{(i)}$ in time t; and $\neg a_j^{(i)}$:= there is no frame buffered at the AP for $s_j^{(i)}$ in time t. Also, let $p_j^{(i)}$ be the probability that, after time t, $s_j^{(i)}$ wakes up unnecessarily, meaning that $s_j^{(i)}$ has to wake up to receive a TIM beacon while there is no frame buffered for it. Using the superposition property of the Poisson process and the independence of the arrival process, we can derive the following equations.

$$\begin{aligned} p_j^{(i)} &= Pr(A^{(i)})Pr(\neg a_j^{(i)}|A^{(i)}) && (4.1)\\ &= Pr(\neg a_j^{(i)} \cap A^{(i)}) && (4.2)\\ &= Pr(\neg a_j^{(i)})Pr(A^{(i)}_{-s_j^{(i)}}), && (4.3)\end{aligned}$$

where $A^{(i)}_{-s_j^{(i)}}$ is an event that $g^{(i)}$ is paged when $s_j^{(i)}$ is excluded from the group. Given that the number of sensory STAs in a TIM group is large and the their individual traffic rates are small, the right hand side of Eq. (4.3) approaches to $Pr(A^{(i)})$. This implies that with high probability a sensory STA will see no buffered frame for it whenever it has to wake up and listen to the TIM

[2] Since the majority of the devices are sensory devices by the assumption, it does not make any notable difference whether or not we take the controllable devices into consideration for the following calculation.

beacon. That is, in the event of $A^{(i)}$, a sensory STA is highly likely to waste energy because of its unnecessary wake-up. Based on this observation, we get the expected amount of energy waste in time t for sensory STAs in $g^{(i)}$ as: $Pr(A^{(i)}) \cdot n_s^{(i)} \cdot \eta$, where η is the amount of energy waste from an unnecessary wake-up, including the energy waste from making state transitions, receiving/processing a TIM beacon and staying in active state. Note that we only consider the energy waste from sensory STAs because the majority of STAs are sensory by assumption. In this regard, we conclude that the expected energy waste from a group is proportional to the total arrival rate to the group, and it is largely dependent upon the sum rate of controllable STAs whose arrival rates are relatively larger than that of sensory STAs.

Therefore, providing an additional membership to controllable STAs, not sensory, and allowing them to dynamically switch to another group will change the expected energy waste of the entire network to a large extent. In what follows, we propose the optimal secondary association program which selectively assigns the unused AID slots to some controllable STAs so as to minimize the number of unnecessary wake-ups.

4.5.3 Problem Formulation

The secondary AID assignment problem in P. 4.4 finds the best strategy $\mathbf{X}$ that allows some controllable STAs to make additional associations, and by which the number of sensory STAs that wake up for nothing is minimized. To this end, we have set a counter-intuitive objective function which is essentially to maximize the expected energy waste instead of minimizing it. Note that the proposed method provides an alternative or extra membership to some STAs (i.e., securing a freedom of switching between groups for some STAs), and leverages it in an opportunistic manner, while leaving the primary association as it is. The key idea of the proposed method is to opportunistically merge buffered traffic to a smaller number of groups so that the number of unnecessary wake-up events to sensory STAs is minimized. The opposite approach, e.g., balancing the sum traffic rate over TIM groups, does not help to reduce the number of unnecessary wake-ups.

The optimization problem P. 4.4 searches for the best secondary association strategy by which the overall expected energy waste is maximized. As a result, each group has a higher chance of turning other paged TIM groups into sleep by redirecting their traffic to itself, and serving them during its own TIM interval.

$$\max_{\mathbf{X}} . \sum_{i=1}^{M} \delta^{(i)} \tag{4.4a}$$

subject to:

$$X_{j,(k)}^{(i)} \in \{0,1\}, \ \forall i, j, k \tag{4.4b}$$

$$\sum_{\substack{i=1,\\ i\neq k}}^{M} \sum_{j=1}^{n_c^{(i)}} X_{j,(k)}^{(i)} \leq y^{(k)}, \ \forall k \tag{4.4c}$$

$$\sum_{k=1}^{M} X_{j,(k)}^{(i)} = 1, \ \forall i, j \tag{4.4d}$$

In the objective function Eq. (4.4a), the term $\delta^{(i)}$ indicates the increase in the expected amount of energy waste due to the associated controllable STAs in TIM group i, which is explained in detail as follows. As an extension of the previously defined event $A^{(i)}$, let us define the following two more events, $A_{+c}^{(i)}$ and $A_{-c}^{(i)}$. The event $A_{-c}^{(i)}$ represents the case that there is at least one frame pending at AP for TIM group $g^{(i)}$ when there is no controllable STA in the group. Similarly, $A_{+c}^{(i)}$ represents the same event except that it is for the case when the group has some controllable STAs, and which controllable STA to associate with is determined by the decision variable $\mathbf{X} = \{X_{j,(k)}^{(i)}\}$. In a mathematical expression, $\delta^{(i)}$ is defined as:

$$\begin{aligned} \delta^{(i)} &= [Pr(A_{+c}^{(i)}) - Pr(A_{-c}^{(i)})] \cdot n_s^{(i)} \cdot \eta \\ &= e^{-\sum_{j=1}^{n_s^{(i)}} \lambda_j^{(i)} t}(1 - e^{-\sum_{k=1}^{M} \sum_{j=1}^{n_c^{(k)}} \mu_j^{(k)} X_{j,(i)}^{(k)} t}) \cdot n_s^{(i)} \cdot \eta. \end{aligned}$$

The decision variable $X_{j,(k)}^{(i)}$ is binary by the constraint Eq. (4.4b), and it is 1 when a controllable STA j which already made a primary association with TIM group i is allowed to make a secondary association with TIM group k, and 0 otherwise. In this regard, $\delta^{(i)}$ measures how much more energy will be wasted if a TIM group $g^{(i)}$ allows controllable STAs to associate with the group compared to the case when the group does not allow any.

The second constraint Eq. (4.4c) indicates that, for each TIM group $g^{(k)}$, the number of secondary associations allowed should not to exceed the number of unused AID slots in the group. The reason why the index k is excluded at the outer summation is that if a controllable STA that belongs to $g^{(k)}$ by the primary association happens to associate with the same group again through the secondary association, it is regarded as the STA is not allowed to make a secondary association and thus, the STA does not actually occupy/consume any AID slot. At last, each controllable device is allowed to make only one secondary association by the constraint Eq. (4.4d).

4.5.4 Relaxation on Binary Constraints

The optimization problem P. 4.4 is not convex, in general, because of the combinatorial nature of the secondary AID assignment whose complexity increases exponentially with the number of controllable STAs and the number of groups [114]. The problem as it is might be solved in a reasonable amount of time if the number of the binary variables is small, but that is not the case for a large-scale 802.11ah network where an AP can associate with up to ∼8,000 STAs. In order to transform P. 4.4 into a tractable, convex optimization problem, the binary constraint in Eq. (4.4b) is relaxed into Eq. (4.5b) by letting each binary variable take any value in [0,1].

$$\max_{\mathbf{X}}. \sum_{i=1}^{M} \delta^{(i)} \tag{4.5a}$$

subject to:

$$X_{j,(k)}^{(i)} \in [0,1], \ \forall i, j, k \tag{4.5b}$$

constraints in (4.4c), (4.4d)

Due to the relaxation of the binary variable, however, $\mathbf{X}$ no longer tells a clear membership relation since it is highly likely to be a real number between 0 and 1, not a binary number. Still, the relaxed decision variables can be used in practice, for example, by interpreting them as either of the following two cases. One is to let $X_{j,(k)}^{(i)}$ be the time sharing factor that indicates the proportion of time that STA j in $g^{(i)}$ can be temporarily associated with $g^{(k)}$. The other is to let $X_{j,(k)}^{(i)}$ be the probability that a corresponding association is allowed for each time. Although both interpretations

Algorithm 4 Iterative one-by-one removal

```
1: repeat
2:     Solve the relaxed optimization problem P. 4.5
3:     for each group i do
4:         $(j^*, k^*) = \arg\min_{\forall j,k} X^{(i)}_{j,(k)}$ s.t. $X^{(i)}_{j,(k)} \neq 0$
5:         Set $X^{(i)}_{j^*,(k^*)} = 0$
6:     end for
7: until all $X^{(i)}_{j,(k)}$ are binary
```

are practically feasible, they incur additional operational complexity to AP and energy consumption to devices due to the frequent change of secondary association; in the worst case, all STAs may have nonzero values for all X's.

4.5.5 Recovering the Binary Solution

In order not to cause the aforementioned extra operational complexity and energy consumption, we have applied a technique that iteratively obtains binary values from the non-binary solutions from P. 4.5. Iterative one-by-one removal algorithm [36] [43] [44] [113] recovers binary solutions from the relaxed ones as described in Algo. 4.

The Algo. 4 first solves the relaxed optimization problem P. 4.5. Next, it finds the nonzero minimum $X^{(i)}_{j,(k)}$ for each group. After forcing up to M number of such variables to be zero, the algorithm returns to Step 2 in line 2 and repeats the whole procedure until all decision variables $X^{(i)}_{j,(k)}$ are binary. Note that on each iteration, instead of forcing multiple variables to be zero, forcing one over all i, j and k would yield a better solution in terms of the optimality gap. However, as the number of both controllable STAs and TIM groups increases, the one-variable-per-iteration approach significantly increases the runtime of the procedure. In order to expedite the one-by-one removal procedure, we have introduced the inner fop-loop which iterates over each group and thus, the algorithm sets up to M variables to be zero for each outer iteration; in Section. 4.7.2, we show that the expedited version still yields a small optimality gap.

4.5.6 Notification Procedure

As a result of the Secondary AID Assignment procedure, some controllable STAs are chosen to be given S-AIDs. The chosen STAs need to know which AID slots they can additionally associate with, which can be done by one of the three methods in general, i.e., broadcast, unicast and piggyback [3]. Among those three, piggyback spends the least amount of energy in this case since it minimizes the number of bits to be transmitted. Thus, in order to efficiently notify the selected STAs of the decision on the secondary association (i.e., S-AIDs), an AP piggybacks an S-AID in the first data to be delivered to the corresponding STA. It is worth mentioning that the secondary association is a long-term decision, meaning that the notification is required only a few times for each chosen STAs during their lifetime. Since the number of controllable STAs is small, and the Secondary AID Assignment procedure chooses only some of them depending on the number of available AID slots, the contribution of the notification procedure to the network-wide energy consumption is trivial.

4.5.7 Discussion

According to [9, Section 4.3.1], "an AID can indicate a groups of STAs." Considering that the main cause of the unnecessary wake-up problem is for using a small number indicators (i.e., small number of bits in DTIM) to represent the pending traffic status for many STAs, it cannot be a solution to the unnecessary wake-up problem. Rather, it may worsen the problem because the maximum number of STAs that can be associated with a TIM group increases. However, the single-AID-for-many-STAs can be beneficial to the proposed scheme in this chapter. The proposed method uses unassigned AID slots to make secondary association. In the worst case, if the network is so populated that there is no vacant slots at all, then the proposed scheme cannot make any secondary association.

However, if an AP can group multiple STAs and then, assign a single AID for them, the number of AIDs in use can be reduced. In other words, an AP can increase the number of unused/unassigned

[3] Piggyback can be regarded as a special type of unicast.

AIDs, and thus to secure more freedom in making secondary association. In sum, if the single-AID-for-many-STAs is used, the proposed method can make more secondary association for the increased number of vacant AID slots. The tradeoff is an increased complexity. Since an AID no longer clearly indicates a physical STA, there should be a mechanism to identify which physical STA is referred to when an AID is paged. Since we are focusing on a low-complex, online scheduling scheme in this chapter, we assume the basic mode, i.e., an AID refers to a single physical STA.

4.6 Algorithm Design: Traffic Scheduling

In this section, we propose a traffic scheduling algorithm that minimizes the number of unnecessarily waking up sensory STAs by taking advantage of the secondary associations made by the Secondary AID Assignment procedure. On each DTIM interval, an 802.11ah-compliant AP constructs a traffic indication map for each group by using P-AID; let us denote this TIM message by *default* TIM. In the proposed scheme, the AP also checks if it can reduce the number of unnecessary wake-ups by using the secondary associations without missing any data to deliver. If it can, the AP rearranges traffic delivery; if not, the AP uses the default TIM as it is.

4.6.1 Exhaustive Search

Intuitively, the optimal traffic scheduling strategy can be found by exhaustive search. It schedules the buffered traffic in all possible ways and chooses the best one that minimizes the number of unnecessary wake-ups. To be specific, the exhaustive search algorithm first enumerates all possible permutations out of M TIM groups. Let us denote $\mathcal{P}$ as the set of permutations which has $M!$ elements, where the i^{th} element $p_{[i]} = \{g_{[1]}, g_{[2]}, \cdots, g_{[M]}\}$ is an ordered list of M TIM groups. Note that $g_{[1]}$ is the first element in the list $p_{[i]}$ and it does not necessarily need to be $g^{(1)}$. On the i^{th} iteration, the AP takes a list $p_{[i]}$ from $\mathcal{P}$. Starting from $g_{[1]}$, the AP rearranges the traffic delivery such that as many P-AIDs and S-AIDs in the TIM group can be utilized as possible if there is buffered traffic associated with those. Then it moves on to $g_{[2]}$ and so on up to $g_{[M]}$. After iterating over all elements in $\mathcal{P}$, it chooses the one with the least number of unnecessary wake-ups.

4.6.2 The Proposed Fast Traffic Scheduling

Although the exhaustive search guarantees to yield the optimal scheduling strategy, it is not suitable for an on-line processing for a large-scale network due to the high computational complexity. Also, it consumes a large memory space because it has to temporarily store all possible permutations of entire groups. In this regard, we propose a lightweight, fast traffic scheduling algorithm that yields a comparable performance to the exhaustive search without causing the aforementioned overhead. The proposed traffic scheduling algorithm is composed of three steps in sequence: 1) Feasibility Check, 2) Preprocessing, and 3) Traffic Scheduling.

4.6.2.1 Feasibility Check

The proposed algorithm first checks if it is possible to reduce the number of unnecessary wake-ups by examining the paged status of sensory STAs. Since sensory STAs are not able to change their membership, having a buffered data for a sensory STA implies that the group to which the STA belongs cannot help but waking up so as to listen to the TIM message. If every group has at least one paged sensory STA, there is no room for improvement. Therefore, the proposed algorithm immediately terminates, and the AP uses the default DTIM and TIM as they are for delivering the buffered frames. On the other hand, if there is at least one group that does not have any paged sensory STAs, there might be a chance that the group can stay in sleep even if some controllable STAs in the group are paged. In such a case, the proposed algorithm proceeds to the next Preprocessing procedure.

4.6.2.2 Preprocessing

The main purpose of having the preprocessing stage is to reduce the complexity of the following Traffic Scheduling procedure. As aforementioned, a TIM group with at least one paged sensory STA must wake up. Therefore, letting paged controllable STAs which made secondary association with such a must-wake-up group switch to the group does not increase (or may decrease) the number of overall unnecessary wake-ups. As a result, the must-wake-up TIM groups schedule as much traffic as possible by using both P-AID and S-AID, and then their scheduling is finalized. Now, the remaining groups are either of the following two cases: 1) that do not have any paged STAs, or 2) that only have one or more paged controllable STAs. For groups that belong to the former case, finalize their schedule as they are. Finally, the remaining groups are the ones that belong to the later case. If the number of the remaining groups is less than or equal to 1, terminate the procedure and return the updated traffic delivery schedule; otherwise, proceed to the Traffic Scheduling procedure.

4.6.2.3 Traffic Scheduling

This stage determines a traffic delivery schedule that minimizes the number of unnecessary wake-ups by leveraging the secondary association which allows some STAs to temporarily change their membership. Given that the remaining groups are the ones that have buffered traffic at AP only for controllable STAs, the traffic scheduling procedure runs as follows. First, it calculates the cost $c(i)$ for all remaining groups by using Eq. (4.6):

$$c(i) = \frac{1}{n^p(i)}(\alpha \cdot \check{n}_c^p(i) + \check{n}_c^s(i)), \tag{4.6}$$

where i is the group index, $n^p(i)$ is the number of STAs associated with group i through P-AID (i.e., $n^p(i)$ is the number of STAs to wake up if the group is paged), $\check{n}_c^p(i)$ is the number of paged controllable STAs associated with group i through P-AID, $\check{n}_c^s(i)$ is the number of paged controllable STAs associated with group i through S-AID, and α is a positive weight to $\check{n}_c^p(i)$. Next, the algorithm selects the group with the maximum cost. If there are multiple groups with the same, maximum cost, the one with the smallest index will be chosen. The algorithm, then, schedules as much traffic as possible to the chosen group by utilizing both P-AID and S-AID. These steps iterate until there is no more group to schedule. According to the Eq. (4.6), the group that has a smaller number of STAs to wake up (if paged), and also has a larger number of paged STAs will be chosen first for having a largeer cost value. In addition, the design parameter α is set to $1 + 1e - 10$ so as to further reduce the number of wake-ups by assigning a slightly larger weight to the controllable devices with P-AID than those with S-AID.

Algorithm 5 Proposed Traffic Scheduling Algorithm

```
Input: $\mathcal{G}, \mathcal{T}$                    // TIM group, default DTIM/TIM info.
// 1) feasibility check:
if every group has paged sensory STA then
    Return $\mathcal{T}$ and terminate
end if
// 2) preprocessing:
for each group $g^{(i)} \in \mathcal{G}$ do
    if $g^{(i)}$ has paged sensory STA then
        Schedule as much traffic as possible to S-AID in $g^{(i)}$
        Update $\mathcal{T}$
        $\mathcal{G} \leftarrow \mathcal{G} - \{g^{(i)}\}$
    end if
end for
if $|\mathcal{G}| \leq 1$ then
    Return $\mathcal{T}$ and terminate
end if
// 3) traffic scheduling:
repeat
    Calculate the cost $c(i)$ for $\forall g^{(i)} \in \mathcal{G}$
    $i^* = \min\{\arg\max_i c(i)\}$
    Schedule as much traffic as possible to S-AID in $g^{(i^*)}$
    Update $\mathcal{T}$
    $\mathcal{G} \leftarrow \mathcal{G} - \{g^{(i^*)}\}$
until $\mathcal{G}$ is empty
Return $\mathcal{T}$ and terminate
```

The overall procedure of the proposed scheduling algorithm is shown in Algo. 5, where $\mathcal{T}$ is the set of the default DTIM and all TIM messages. Please note that the traffic delivery for those groups that are not mentioned in the algorithm will be made as specified by the default TIM. Compared to the exhaustive search method of which search space (i.e., the number of groups to be explored) is $\Theta(M!)$, the proposed scheduling algorithm reduces it down to O(M).

It is worth mentioning that the expected amount of additional delay caused by the proposed method is zero compared to 802.11ah. On each DTIM interval, the proposed scheme delivers exactly the same amount of traffic that 802.11ah has scheduled to. In addition, the expected amount of delay within a DTIM interval is also zero. Let $d^{(i)}$ be such a delay that STAs in $g^{(i)}$ experience on traffic delivery, which is defined as:

$$d^{(i)} = \sum_{j=1}^{n_c^{(i)}} \sum_{k=1}^{M} Pr(X_{j,(k)}^{(i)} = 1)\tau_{i,k}, \tag{4.7}$$

where $\tau_{j,(k)}^{(i)}$ is an increase/decrease in time to traffic delivery when a STA j in $g^{(i)}$ by P-AID is served in $g^{(k)}$. We have $\tau_{j,(k)}^{(i)} > 0$ if $g^{(k)}$ is served after $g^{(i)}$, $\tau_{j,(k)}^{(i)} < 0$ if $g^{(k)}$ is served before $g^{(i)}$, and $\tau_{j,(k)}^{(i)} = 0$ if $k = i$. Note that sensory STAs are not allowed to change their membership and thus, they do not contribute to $d^{(i)}$ at all. Since we do not make any assumption on how the primary association is made, an event $X_{j,(k)}^{(i)} = 1$ is regarded to be equally likely among all possible $k = 1, 2, \cdots, M$. Then, the expected amount of delay increase cancels that of decrease and therefore, Eq. (4.7) becomes zero. That is, the expected delay caused by the proposed method is zero.

4.7 Evaluation

We have implemented both the optimal Secondary AID Assignment problem and the Traffic Scheduling algorithm on CVX [50] and MATLAB [49], and compared the performance of the proposed scheme to 802.11ah and the exhaustive search method.

4.7.1 Network Configuration

Evaluation has been performed with various number of TIM groups, each of which can associate with up to 64 STAs [8]. Also, we consider TIM groups in a single page, which can associate with up to 2048 STAs in total. Due to the advanced antenna technology, an AP can interact with STAs in one page independently from the rest, for example, by using the sectorized beam operations [9]. Therefore, the results to be shown in this section should remain the same whether or not we consider multiple pages at the same time. The traffic arrivals follow the Poisson distribution whose mean value is randomly chosen. Also, the primary association is randomly made and thus, each group is highly likely to have different number of STAs and different sum traffic rate from the rest. Each data point on the figures in this section is an average of several runs of simulation, except

Fig. 4.9. To be specific, we have run 10 simulations for both Fig. 4.3 and Fig. 4.4. Also, for Fig. 4.5, Fig. 4.6, Fig. 4.7 and Fig. 4.8, we have run 100 simulations. For each data point in those figures, we have marked a 95% confidence interval. Please note that we have omitted the result of the case when $M = 1$, since the scenario is too simple and it does not allow any secondary associations to make; in other words, there will be no performance difference between the proposed scheme and 802.11ah.

4.7.2 Secondary AID Assignment & Optimality Gap

As discussed in Section 4.5, we have applied relax-and-recover approach (Algo. 4) to the Secondary AID Assignment problem in order to maintain the overall complexity low and to make the initial problem (P. 4.4) tractable. Considering that the objective value of the original problem P. 4.4 is upper-bounded by the relaxed problem P. 4.5, and lower-bounded by the recovered binary solution, we first show that the optimality gap between the two bounds is tight, implying that the recovered binary solution has a high accuracy.

The Fig. 4.3 shows how the objective value changes as the number of TIM groups increases. In the figure, *Relaxed* and *One-by-one removal* correspond to the solution of the relaxed association problem P. 4.5 and the recovered binary solution from running Algo. 4, respectively. Although the objective value of the relaxed problem is always larger than that of the recovered solution, the difference is small, meaning that the recovered solution has high accuracy. Also, we have measured the relative error between the two solutions as shown in Fig. 4.4. The y-axis denotes the ratio of difference between the two objective values to the objective value of the relaxed problem. Throughout all scenarios with different number of TIM groups, the mean error ratio never exceeds 0.005. That is, the bounds on the initial problem P. 4.4 are small and thus, the recovered solution is close to the optimum.

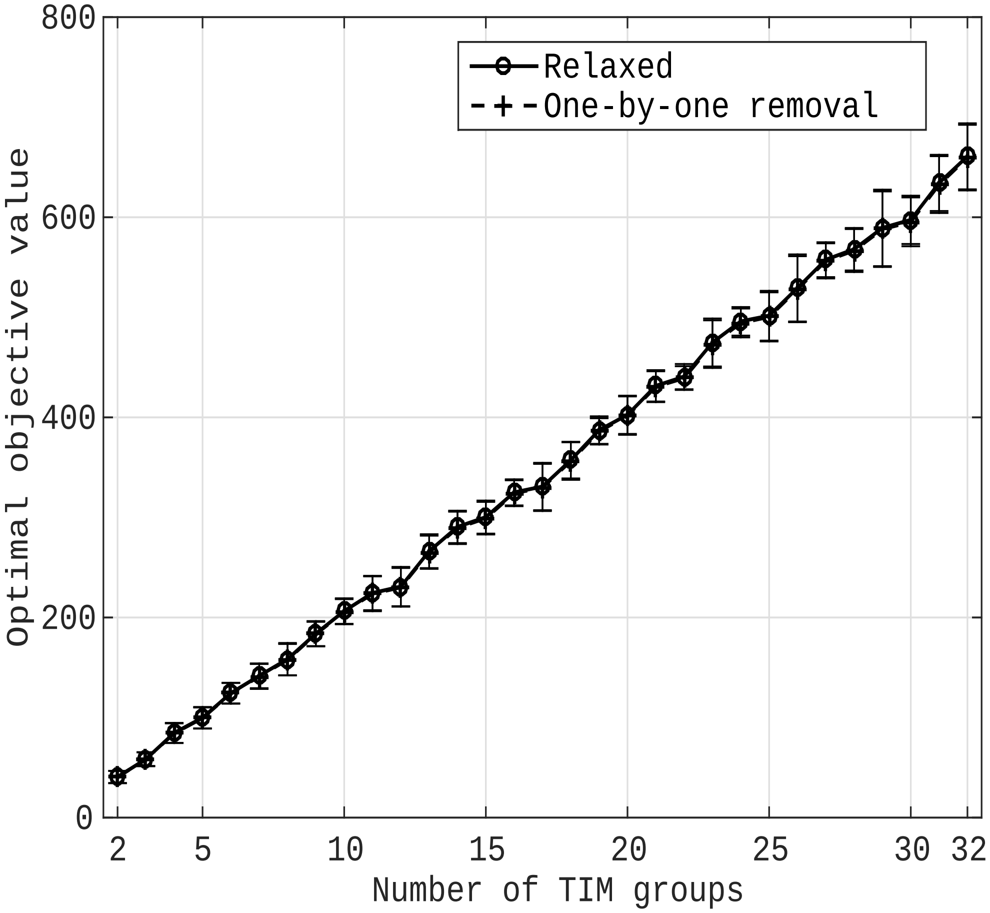

Figure 4.3: Comparison of the optimal objective values between the relaxed problem solution and the recovered binary solution.

4.7.3 Traffic Scheduling & Number of Unnecessary Wake-ups

Next, we introduce the simulation results to show how effective the traffic scheduling algorithm is in terms of the number of unnecessary wake-ups. We compare the number of such energy-wasting wake-ups between 802.11ah, exhaustive search and the proposed method. Since the three methods deliver exactly the same amount of traffic on each DTIM interval, having a smaller number of wake-ups yields a better energy efficiency.

Both Fig.4.5 and Fig. 4.7 show the mean number of total unnecessary wake-ups with different number of TIM groups. Note that due to the high complexity of the exhaustive search method

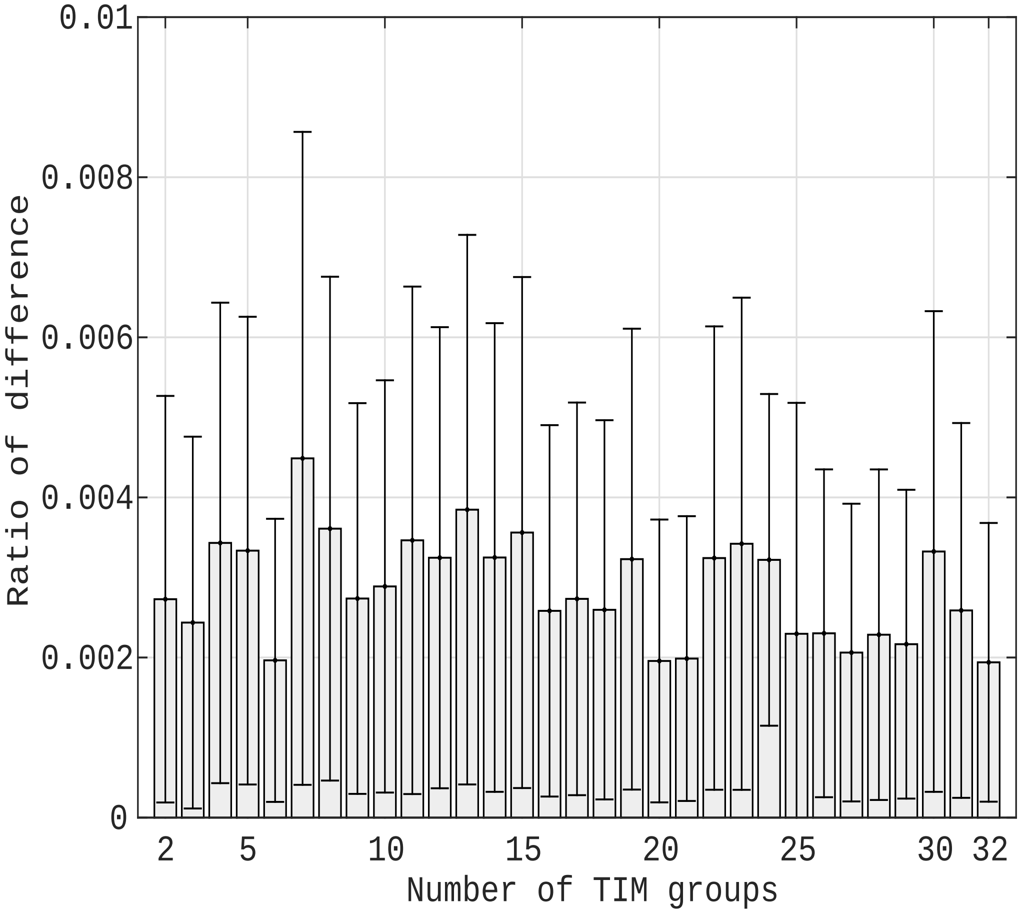

Figure 4.4: The ratio of the difference between the objective values of the relaxed and the recovered binary solution to the objective value of the relaxed solution.

in terms of time and memory space, we were able to run it with up to 10 TIM groups of which results are shown in Fig.4.5. The proposed method, on the other hand, has a short response time and requires only a small memory space. Thus, we have carried out all possible scenarios, i.e., with up to 32 TIM groups, as shown in Fig. 4.7. As it can be seen in Fig.4.5, both exhaustive search and the proposed algorithm outperform 802.11ah in terms of the number of unnecessary wake-ups. As the number of both TIM groups and STAs increases, a larger freedom in making secondary association is exploited and thus, the performance enhancement compared to 802.11ah becomes larger for both exhaustive search and the proposed algorithm. It is noteworthy that the

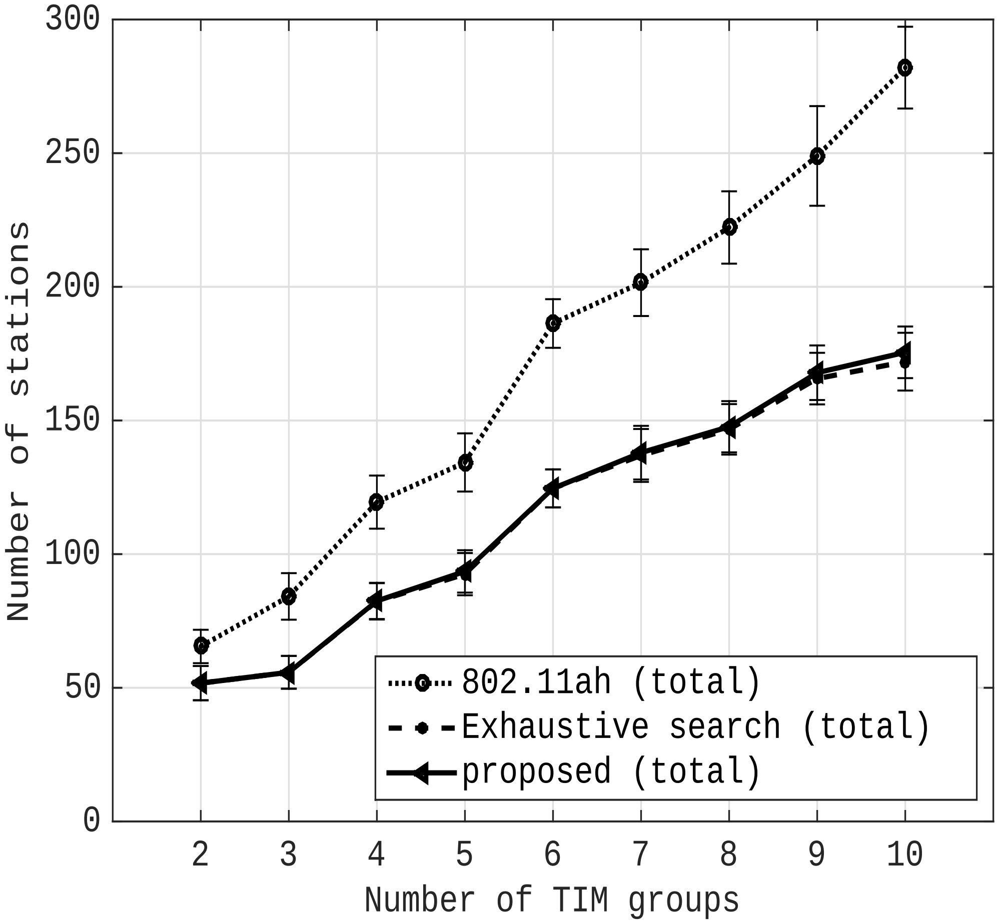

Figure 4.5: The mean number of unnecessarily waking-up STAs per DTIM interval with different number of TIM groups.

proposed scheme yields a comparable performance to the exhaustive search solution, although it has a much smaller computational and resource overhead. In addition, we have counted the number of unnecessary wake-ups for sensory STAs separates from that for controllable STAs as shown in Fig. 4.6. Since there are more sensory STAs than controllable by assumption, three lines with larger values than the rest indicate the number of unnecessary wake-ups for sensory STAs, while the three with lower values are for controllable STAs. As it can be seen in the figure, 802.11ah produces the largest number of unnecessary wake-ups for both sensory and controllable STAs. Although the Secondary AID Assignment problem focuses on minimizing energy waste from sensory STAs, both

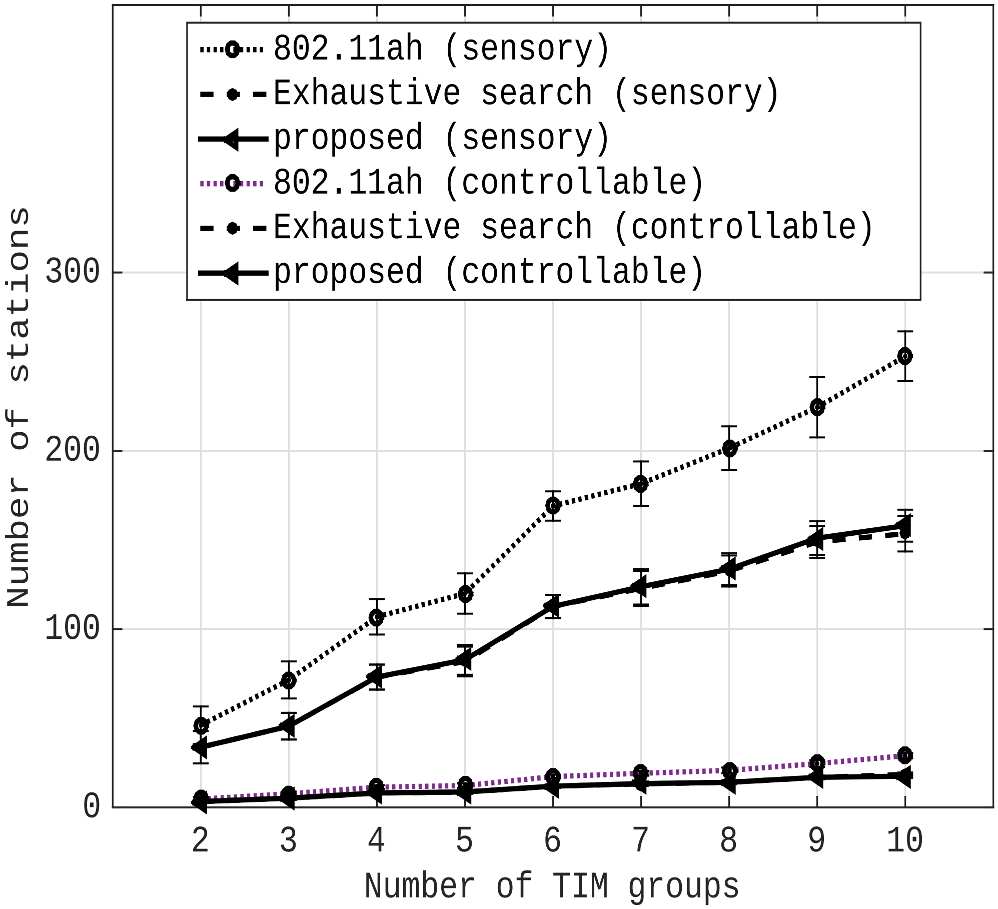

Figure 4.6: The mean number of unnecessarily waking-up sensory and controllable STAs per DTIM interval with different number of TIM groups.

the proposed method and the exhaustive search still outperform 802.11ah in terms of the energy waste from controllable STAs as well.

The following Fig. 4.7 shows the number of unnecessary wake-ups for both 802.11ah and the proposed method as the number of TIM groups increases up to 32. Again, we have separately shown the unnecessary wake-ups for sensory and controllable STAs in Fig. 4.8. As the number of both TIM groups and STAs increases, the difference in the number of unnecessary wake-ups becomes larger between 802.11ah and the proposed scheme, meaning that the proposed scheme

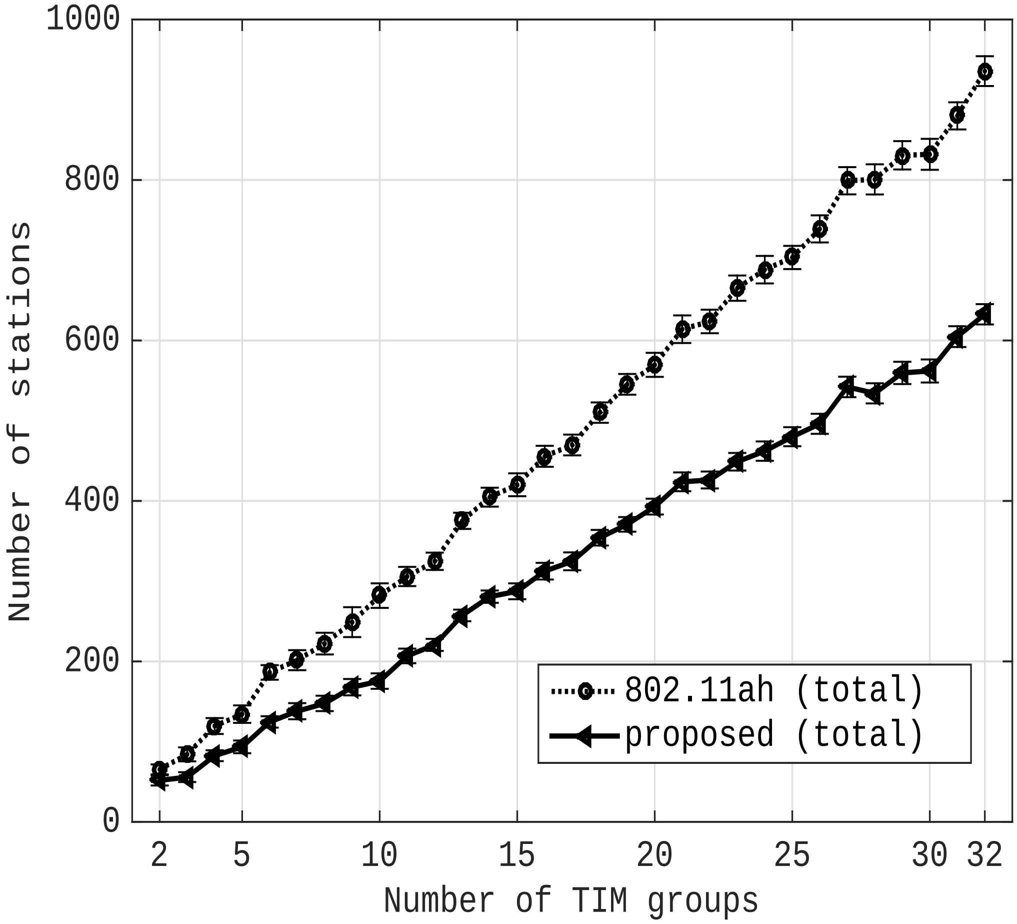

Figure 4.7: The mean number of unnecessarily waking-up STAs per DTIM interval with different number of TIM groups.

brings more gains to enhancing energy efficiency. For instance, when there are 32 TIM groups, the proposed scheme, on average, incurs 632.6 unnecessary wake-ups, while 802.11ah does 935.7.

The performance enhancement in [0,1] scale can be measured by the ratio of the difference in the number of unnecessary wake-ups between 802.11ah and the proposed scheme to that of 802.11ah. Except the case of $M = 2$ which results in a ratio of 0.210, the performance enhancement is between 0.303 and 0.378 for the rest of M values, meaning that at least 30% of more devices are able to stay in low-power state when the proposed scheme comes into play compared to the 802.11ah power

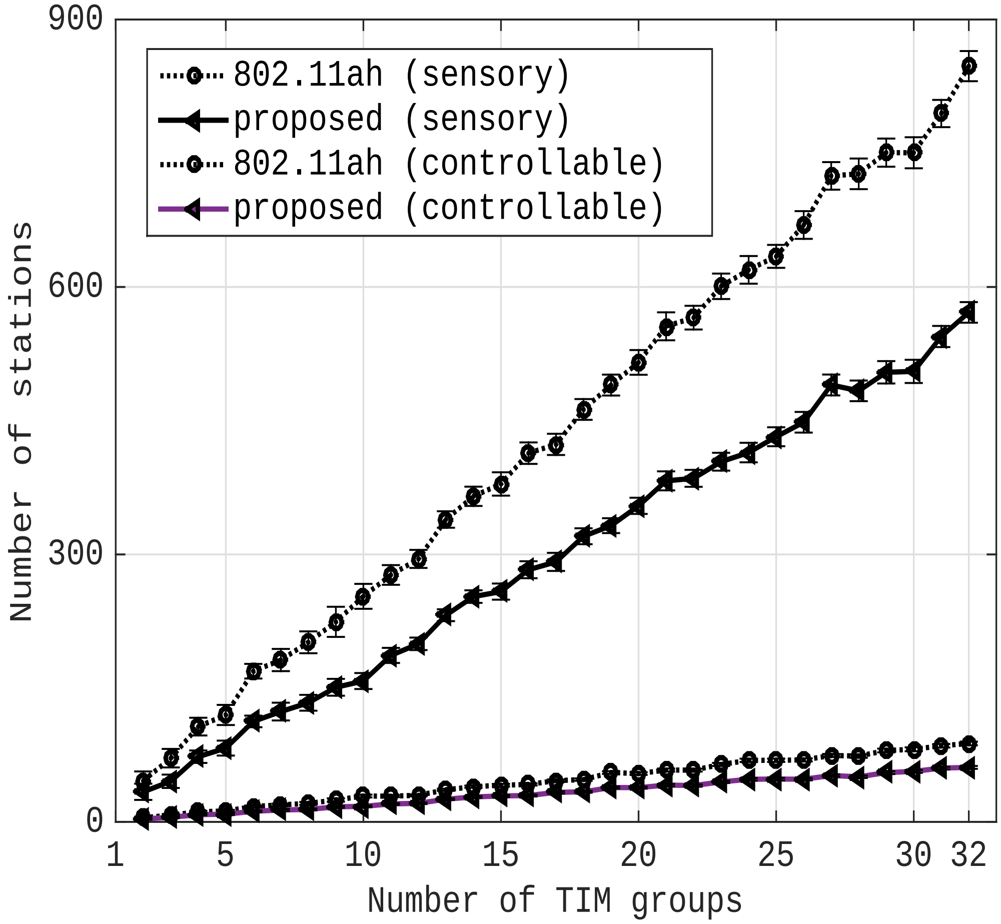

Figure 4.8: The mean number of unnecessarily waking-up sensory and controllable STAs per DTIM interval with different number of TIM groups.

saving mechanism. As a reminder, both schemes deliver exactly the same amount of traffic, and the proposed scheme causes, on average, zero delay compared to 802.11ah.

As studied in [115], the basic power model for power consumption of the receiving circuitry can be modeled as a constant. Therefore, the amount of energy waste due to the unnecessarily waking-up STAs can be simply approximated by a linear function of the number of unnecessary wake-ups, in addition to the constant power usage for staying in active state. Considering the expected lifetime of 802.11ah STAs, and the number of DTIM intervals that STAs are expected to

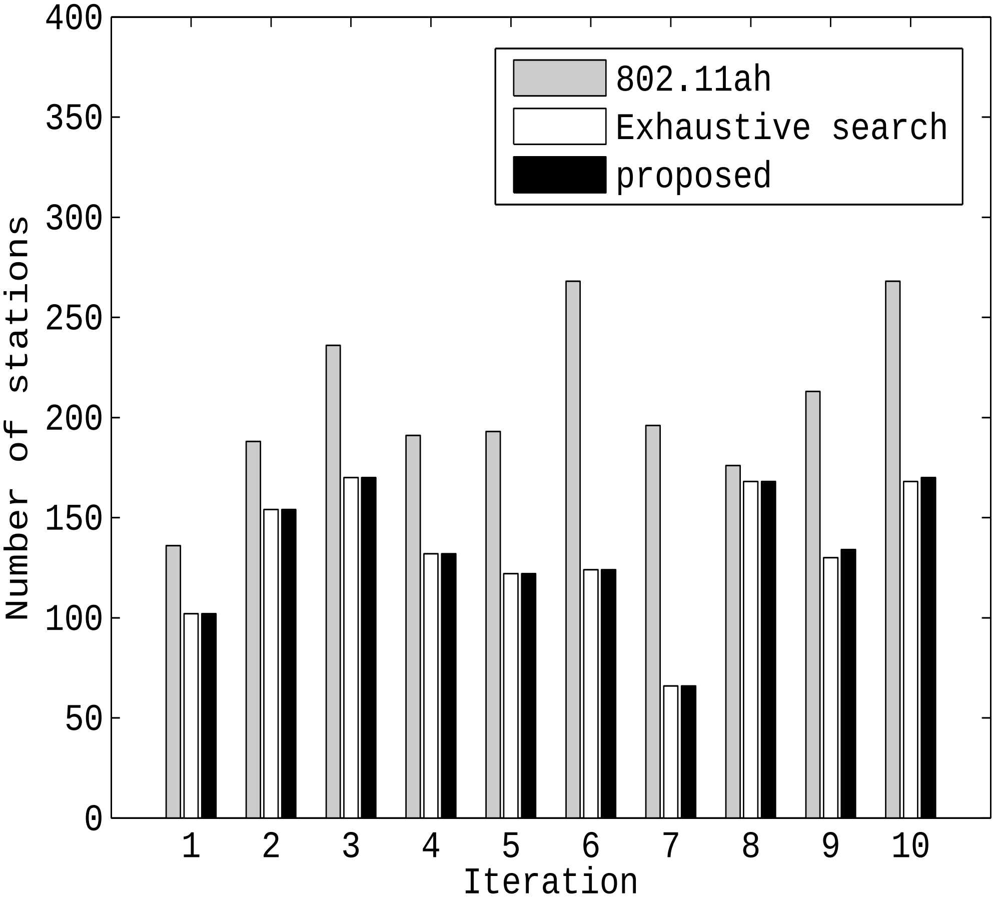

Figure 4.9: Example: the number of unnecessarily waking-up STAs during the first 10 iterations when the number of TIM groups is 8.

encounter during operation, reducing the number of unnecessary wake-ups will significantly prolong the lifetime of STAs as well as the whole network.

Yet, the computational and memory efficiency of the proposed scheme comes at the expense of the performance degradation compared to the exhaustive search method as shown in Fig. 4.9. On iterations 9 and 10, exhaustive search outperforms the proposed method in terms of the number of unnecessary wake-ups. However, the amount of performance degradation is trivial, and it becomes negligible as the iteration goes on as shown in Fig. 4.5.

4.8 Conclusion

In this chapter, we have studied a large-scale IEEE 802.11ah wireless sensor network. In particular, we have focused on the energy efficiency of power saving mechanism on top of TIM and page segmentation. To reduce the power consumption of battery-operated sensor devices, we have proposed a novel way of utilizing unused AIDs and scheduling traffic delivery so as to reduce the number of unnecessary wake-ups. The proposed temporary membership change scheme allows the traffic scheduling algorithm to exploit a certain level of freedom in delivering buffered traffic and thus, the number of unnecessary wake-ups is minimized. In order to expedite the entire procedure, the relaxation and recovery scheme is applied to the optimization program, and a lightweight traffic scheduling algorithm is proposed. The evaluation results show that the proposed algorithm can significantly increase the number of sensor devices that can stay in low-power state, which could prolong their lifetime without increasing delay to traffic delivery.

CHAPTER 5. ENERGY-EFFICIENT SCHEDULING OF INTERNET OF THINGS DEVICES FOR NEXT GENERATION WLANS: ENVIRONMENT MONITORING APPLICATIONS

5.1 Summary

The Internet of Things (IoT) paradigm has been proposed to assist and automate various activities such as environment monitoring by connecting physical devices in the area of our interest. Low-cost, battery-operated, and resource-limited IoT devices usually are densely deployed for robustness against node failures as well as for providing desired quality of monitoring. One of the effective ways to prolong the lifetime of an IoT network is to schedule selected IoT devices to enter the sleep mode and activate them later in a future time. However, this must be done carefully in order not to violate application-specific requirements. In this chapter, we study the scheduling problem of IoT devices to prolong the network lifetime while satisfying both report error/accuracy[1] and timely update (or reporting interval) requirements. We model and analyze the network as a Markov process, and derive system parameters. We formulate an optimal node activation scheduling problem and propose a low-complex greedy algorithm to expedite the scheduling process. Evaluation results demonstrate the effectiveness of both optimal and greedy algorithms. The complete version of this chapter has been published in [116].

5.2 Introduction

Low-power, low-cost devices with sensing modules have been widely used in our daily lives as well as in special circumstances for particular missions such as environment monitoring and target tracking. Despite such devices' limited computing power and energy supply, they have become the essential component for Internet of Things (IoT) and Wireless Sensor Networks (WSNs). Such

[1] In this study, we use the report accuracy requirement and the report error requirement interchangeably.

trend is expected to prosper by IEEE 802.11ah TG (Task Group) [9] that is intended to provide a unified wireless communication solution for large-scale IoT networks. To be specific, an 802.11ah-compatible AP (Access Point) is capable of associating with up to 8,192 devices with a coverage radius of about 1 km [85].

Although 802.11ah has built-in power saving methods [85], energy efficiency is still one of the biggest challenges for many IoT applications. This is because IoT devices (referred to as *nodes*) are battery-operated, and it may not be cost-effective or even possible to recharge or replace batteries in some applications, e.g., environment monitoring of an inaccessible terrain. In general, nodes are densely deployed in an IoT network for connectivity and coverage as well as for dealing with nodes' high failure rates [117]. Highly-populated nodes, however, increase the channel contention for the shared medium as well as produce redundant sensed readings, which decreases the overall energy efficiency and the network lifetime.

One effective way to prolong the network lifetime is to operate some of the nodes in the *sleep mode* and activate them later in a future time. However, the reduced number of active nodes may degrade the system performance, such as the accuracy of the fusion report in environment monitoring applications. Also, the reduced number of active nodes may result in a violation of some application-specific requirements, such as maximum update interval or maximum delay requirement in environment monitoring applications.

In this chapter, we study the problem of energy-efficient activation/sleep scheduling of IoT devices to maximize the network lifetime for environment monitoring applications. Here, the network lifetime refers to the time period during which the network operates properly without violating any application-specific requirements. As shown in Fig. 5.2, we consider a clustered IoT network with multiple fusion centers (FCs) serving as cluster heads to collect sensed readings from IoT nodes and then transmit the fusion reports to the AP. We consider the following two requirements: (1) [report error/accuracy requirement] maximum error in the fusion report created by the FC, assuming that each sensed reading has an inherent measurement error; and (2) [timely update (or reporting interval) requirement] maximum reporting interval from the FC to the AP.

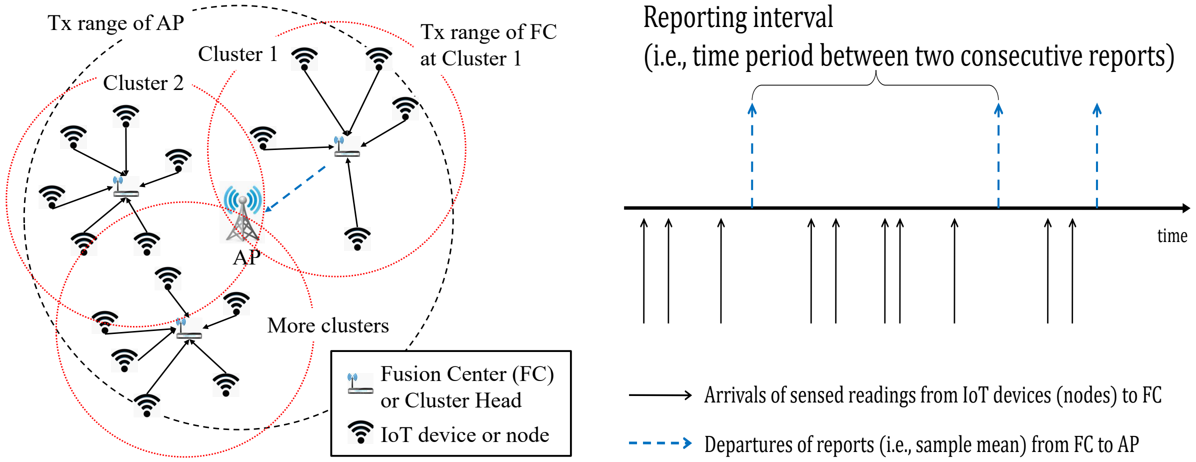
Tx range of AP
Cluster 1
Cluster 2
Tx range of FC
at Cluster 1
AP
More clusters
Fusion Center (FC)
or Cluster Head
IoT device or node
Reporting interval
(i.e., time period between two consecutive reports)
time
Arrivals of sensed readings from IoT devices (nodes) to FC
Departures of reports (i.e., sample mean) from FC to AP

Figure 5.2 Network diagram with an AP, multiple clusters, and randomly deployed nodes.

Figure 5.3 Illustration of arrivals of sensed readings to an FC (solid lines) and departures of reports from the FC (dashed lines).

Figure 5.4: System model.

To do so, we first model and analyze the network as a Markov process. Then, we compute the optimal number of nodes to activate, and configure the FC to satisfy the requirements. Finally, given the optimal number of nodes to activate, we formulate an optimal node scheduling problem to determine the set of nodes to activate by taking into consideration the nodal residual energy. The problem is formulated as a mixed integer linear program (MILP) problem that is intractable in general; thus, we also propose a low-complexity greedy algorithm to solve the problem.

The rest of this chapter is organized as follows. The network model and problem formulation are given in Section 5.3. We analyze the network in Section 5.4. In Section 5.5, we derive an optimal node scheduling algorithm, and propose a practical greedy algorithm with a lower complexity. Evaluation results are presented in Section 5.6. Finally, we conclude the chapter in Section 5.7.

5.3 System Model and Problem Overview

5.3.1 Models and Assumptions

We study a clustered 802.11ah-compatible IoT network for environment monitoring applications. As shown in Fig. 5.2, many battery-operated and resource-limited IoT devices (referred to as *nodes*) are deployed uniformly at random in the area to be monitored, while multiple FCs are deployed to relay the data from IoT devices to the AP in the form of fusion reports. Each FC has a stable power supply and can communicate directly with the AP. An FC associates with the nodes within its transmission range, forming a cluster. We use $\mathcal{W}$ to denote the set of nodes in a cluster. If a node is within the transmission range of multiple FCs, it associates with the one with the strongest signal strength.

An active node senses the environment (e.g., temperature, sound, light, or vibration) and transmits the sensed readings (referred to as *data*) to the FC. After collecting multiple sensed readings from its associated nodes, the FC takes a sample mean, and transmits only the mean value (referred to as *report*) to the AP. For the small size of the message to be transmitted from a FC to the AP, each FC can minimize both the time to occupy the wireless medium and the probability of having an error in the transmitting signal. In this study, we use *reporting interval* to refer to the time interval between two consecutive reports from the FC to the AP, as illustrated in Fig. 5.3.

5.3.1.1 From IoT Devices to the Fusion Center

We make the following assumptions on the sensed readings generated by IoT devices of the same cluster within the l^{th} reporting interval: (1) they are generated according to a Poisson distribution with a rate of λ which is the aggregate data rate of all nodes ($\mathcal{W}$) in the same cluster; (2) their values are i.i.d. (independent and identically distributed) random variables, with a mean of $\overline{D^l}$ and a variance of σ^2, where σ^2 can be viewed as the sensing/measurement error. We assume that nodes are homogeneous with inexpensive hardware and simple software; thus, they have the same measurement error of σ^2 and the same fixed data generation rate of $\lambda/|\mathcal{W}|$. Each node has a limited amount of energy supply without any energy harvesting or energy replenishment capabilities. We assume that each FC has a buffer of finite size B, in which up to B number of data can be stored. All the received data are stored in the FC's buffer, until the FC takes a sample mean and transmits the fusion report to the AP. If the buffer becomes full, the FC discards any new data arrivals. Once the FC transmits a report to the AP, it flushes its buffer.

5.3.1.2 From the Fusion Center to the AP

We assume that the reports from the FC to the AP also follow a Poisson distribution with a rate of μ. Different from IoT devices, the FC can adjust its reporting rate μ dynamically. We assume that μ is discrete, and it is adjustable between $[\mu_{\min}, \mu_{\max}]$ with a step size of Δ_μ. Clusters are assumed to be independent of each other. Thus, we focus our study on a single cluster in the rest of this chapter. The optimal solution derived from a cluster can be easily applied to the rest of the clusters. In 802.11ah, each node is assigned a unique identifier, called AID (association ID). Nodes are partitioned into different clusters based on their AIDs. The restricted access window (RAW) method – which is one of the most noteworthy features in 802.11ah – provides each cluster with a dedicated time interval to access the channel exclusively. By limiting the number of nodes that are allowed to compete for the shared wireless medium at the same time, the RAW scheme reduces channel contention, while increasing the time period during which nodes can enter the sleep mode or low-power state.

5.3.2 Problem Statement

For each cluster, our **goal** is to schedule the activation of IoT nodes to maximize the lifetime of the cluster. Recall that the lifetime of a cluster refers to the time period during which the cluster operates properly without violating any requirements. Clearly, having too many active nodes may generate redundant sensed readings and deplete their battery fast. On the other hand, having too few active nodes may render the FC's report (i.e., sample mean) less accurate. In addition, even if the optimal number of nodes to activate is known, the schedule of node activations may greatly affect the lifetime of a cluster. In this regard, we formulate an energy-efficient node scheduling problem as follows.

Given:

- the set of nodes $\mathcal{W} = \{w_1, w_2, \cdots, w_{|\mathcal{W}|}\}$ in a cluster,
- the total arrival rate λ of sensed readings,

- the nodal residual energy $\phi(w_a)$, $\forall w_a \in \mathcal{W}$,
- the measurement error σ^2, and
- the buffer size B at the FC,

with the following **constraints** (or application requirements):

- *report error/accuracy requirement:* the error in a fusion report should not exceed the expected error bound σ_{req} with probability $1 - p_e$ at least, and
- *timely update (or reporting interval) requirement:* the reporting interval (on average) should not exceed the maximum expected reporting interval τ_{req},

the **outputs** of the proposed scheme are:

- the optimal reporting rate μ^* of the FC,
- the optimal number of nodes η^* to activate, and
- the optimal node activation schedule,

such that the lifetime of a cluster is maximized. The **terminating condition** of the proposed method is: $|\{w_a : \forall w_a \in \mathcal{W} \text{ such that } \phi(w_a) \geq \phi_{thr}\}| < \eta^*$, where ϕ_{thr} is the minimum residual energy for an IoT node to function normally. That is, when the number of nodes with enough residual energy is smaller than η^*, the proposed scheme terminates since the cluster can no longer satisfy one or both requirements.

5.3.3 Overview of the Proposed Solution

To achieve the goal, we propose a 3-staged solution as illustrated in Fig. 5.5. In the Network Analysis stage, we first model and analyze the network by using a Markov process, and based on which we compute the least number of nodes (η^*) to activate without violating any of the application requirements. Also, we compute the optimal μ^* at which the FC shall report to the AP. Then, in the Optimal Node Scheduling stage, we derive an optimal scheme to schedule the node activation

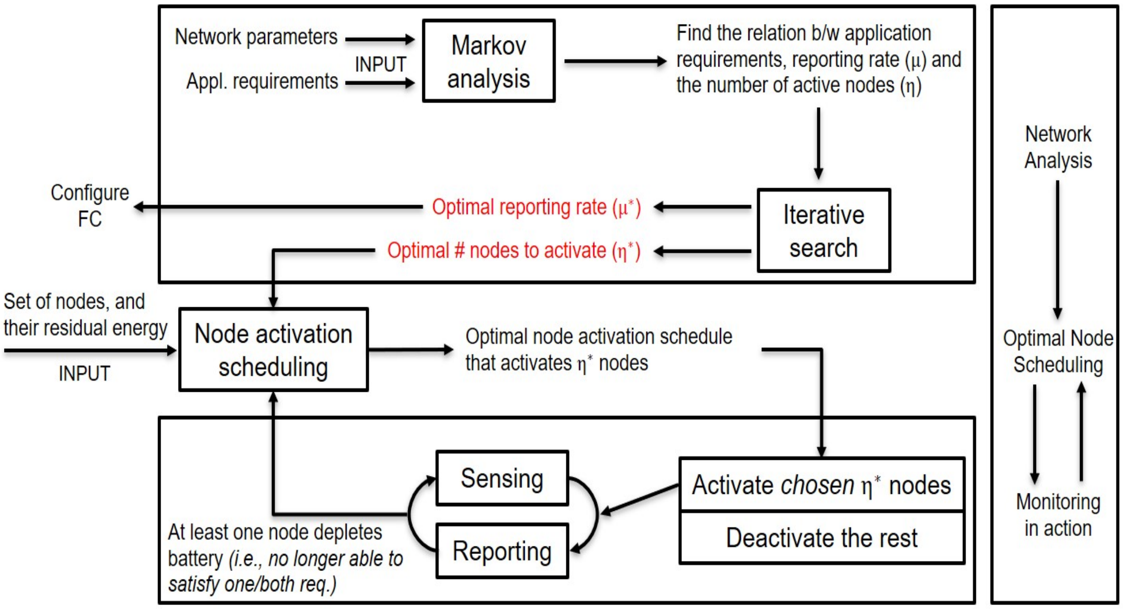
Network parameters
Appl. requirements
INPUT
Markov analysis
Find the relation b/w application requirements, reporting rate (μ) and the number of active nodes (η)
Iterative search
Optimal reporting rate (μ*)
Optimal # nodes to activate (η*)
Configure FC
Set of nodes, and their residual energy
INPUT
Node activation scheduling
Optimal node activation schedule that activates η* nodes
Sensing
Reporting
Activate chosen η* nodes
Deactivate the rest
At least one node depletes battery (i.e., no longer able to satisfy one/both req.)
Network Analysis
Optimal Node Scheduling
Monitoring in action

Figure 5.5: Overview of the proposed solution.

by taking into consideration the residual energy of each node so that we can determine the set of nodes to activate to maximize the cluster lifetime. Afterwards, in the Monitoring stage, the chosen η^* nodes are activated and sensing the environment. Since the proposed scheme activates the optimal (or minimal) number of nodes, even a single failure among the operating nodes (e.g., by depleting the battery) causes the violation of one or both requirements. Then, the proposed solution enters the Optimal Node Scheduling stage again to choose another η^* nodes to activate.

In this study, the way to satisfy the requirements are based on the mathematical analysis, which helps to build a more practical and robust solution than the heuristic ones. By considering the direct relation between the nodal residual energy and the cluster lifetime, the proposed solution performs better than the methods of which solutions are produced based on some indirect relations instead (e.g., maximizing a utility function). In the proposed solution, the FC can adapt its behavior by configuring its reporting rate μ, so that it can adjust to the network dynamics to meet the goal and the requirements. The proposed solution is automated and scalable, and thus is promising for the large-scale IoT applications.

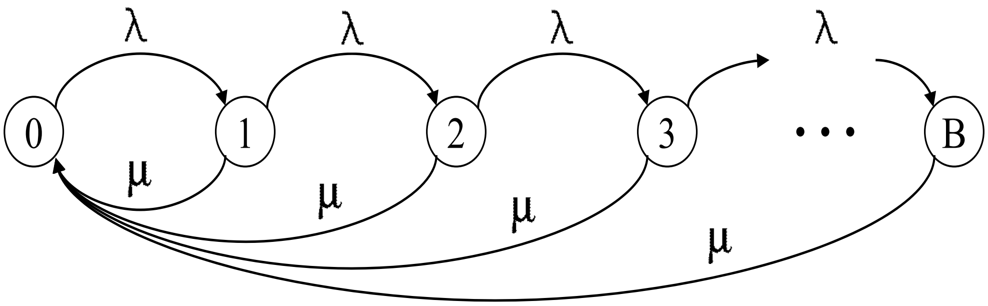

Figure 5.6: Markov chain model for the FC (Fusion Center).

5.4 Network Modeling and Analysis

Let $\{X_t\}_{t\geq 0}$ be a continuous-time stochastic process that takes a value from a countable set $\mathcal{B} = \{0, 1, 2, \cdots, b, \cdots, B\}$ at time $t \geq 0$. Here, $\mathcal{B}$ is a finite set since the FC discards any new arrivals when its buffer is full. By $X_t = b$ (i.e., b is the state of X_t), we mean that the FC has b number of data in its buffer at time t. When $X_t < B$, the state of the process is increased by 1, if there is a new arrival from an active node. Since the shared wireless medium can be occupied by at most one device at a time, the state cannot be increased by more than 1. The state of the process returns to 0 when the FC transmits a report to the AP and flushes it buffer. The amount of time spent in a state (referred to as *holding time*) is continuous, and the next state transition, given the present state, is independent of the past. Thus, $\{X_t\}_{t\geq 0}$ is a continuous-time Markov process with the holding time in each state being exponentially distributed. The corresponding continuous-time Markov chain (CTMC) is shown in Fig. 5.6.

Let $\mathcal{W} = \{w_1, w_2, \cdots\}$ be the set of nodes in a cluster. Arrivals from an active node to the FC is independent of the arrivals from other nodes, and the total rate at which the CTMC transitions to the right is λ, given the current state is not B. On the other hand, the CTMC transitions to state 0 at rate μ, given the current state is not 0; please note that the FC does not report to the AP when it has no data in its buffer. Let $\mathbf{Q} = [q_{ij}]_{(i,j)\in\mathcal{B}\times\mathcal{B}}$ be the transition rate (or infinitesimal generator) matrix where q_{ij} is the rate at which the CTMC transitions from state i to state j.

$$\mathbf{Q} = \begin{bmatrix} -q_{01} & q_{01} & 0 & \dots & 0 \\ q_{10} & -(q_{10}+q_{12}) & q_{12} & \dots & 0 \\ q_{20} & 0 & -(q_{20}+q_{23}) & \dots & 0 \\ \vdots & 0 & 0 & \ddots & \vdots \\ q_{M0} & 0 & 0 & \dots & -q_{M0} \end{bmatrix}$$

Here, we have $q_{ii} = -q_i$ where $q_i = \sum_{\forall j \neq i} q_{ij}$ is the total outgoing rate from state i.

From the CTMC with the rate matrix, we can derive an embedded discrete-time Markov chain (DTMC) with exponential holding time. From DTMC, we know that all the states communicate with each other, forming one communication class, and thus, both DTMC and CTMC are irreducible. An irreducible DTMC with a finite state space is always recurrent (i.e., all states are recurrent), and so is the equivalent CTMC. Also, the CTMC is non-explosive (or regular) since the state space is finite. Since $\{X_t\}_{t\geq 0}$ is irreducible and recurrent, we can compute the stationary distribution $\boldsymbol{\pi} = \{\pi_1, \pi_2, \cdots, \pi_B\}$ by using both the global balance equation (i.e., $\sum_{\forall j \neq i} \pi_i \times q_{ij} = \sum_{\forall j \neq i} \pi_j \times q_{ji}$) and $\sum_{\forall i \in \mathcal{B}} \pi_i = 1$. Let $\rho = \frac{\lambda}{\lambda+\mu}$; then, we can compute π_i for all states i as shown below, where π_i is the proportion of time spent in state i over a long period of time.

$$\pi_i = \begin{cases} \left\{\frac{1-\rho^B}{1-\rho} + \frac{\lambda}{\mu}\rho^{B-1}\right\}^{-1}, & \text{for } i = 0. \\ \rho^i \left\{\frac{1-\rho^B}{1-\rho} + \frac{\lambda}{\mu}\rho^{B-1}\right\}^{-1}, & \text{for } 0 < i < B. \\ \frac{\lambda}{\mu}\rho^{B-1}\left\{\frac{1-\rho^B}{1-\rho} + \frac{\lambda}{\mu}\rho^{B-1}\right\}^{-1}, & \text{for } i = B. \end{cases}$$

Given that $\{X_t\}_{t\geq 0}$ has a unique sum-up-to-1 stationary distribution, the process is positive recurrent.

By using the stationary distribution $\boldsymbol{\pi}$, we can compute the expectation of (1) the number of sensed readings buffered and used by the FC to take the sample mean for a reporting interval, and (2) the length of a reporting interval. The finding in (1) is used to check if the error in a fusion report does not exceed the error bound σ_{req} with probability $1-p_e$ (at least), and the finding in (2) is used to make sure if the reporting interval (on average) does not exceed the maximum expected reporting interval threshold τ_{req}. By combining the two findings together, we can determine both the optimal number of nodes to activate and the optimal reporting rate of a FC to satisfy the two requirements.

5.4.1 Report Error/Accuracy Requirement

Let U^l be a random variable for the number of buffered data that the FC uses to calculate the sample mean within a reporting interval l. For example, if $U^l = b \in \mathcal{B}$, it means that the FC has calculated and reported a sample mean with b sensed readings, and hence CTMC has made a transition from state b to state 0. By using the stationary distribution $\boldsymbol{\pi}$, we can compute the expectation of U^l as: $E[U^l] = \frac{1}{1-\pi_0} \sum_{i=1}^{B} i \times \pi_i$. The summation term is normalized by $1 - \pi_0$ since there is no transition from state 0 to itself. Also, let D^l_m be a random variable for the value of the m^{th} sensed reading that arrives at the FC within a reporting interval l. Since D^l_m's are i.i.d. by assumption, the sample mean for the reporting interval l is a random variable $S^l = \frac{1}{U^l} \sum_{m=1}^{U^l} D^l_m$. By the Central Limit Theorem, S^l follows the Normal distribution with a mean of $\overline{D^l}$ and a variance of σ^2/U^l.

Let us assume that S^l has a finite expected value $E[S^l]$ and a finite non-zero variance $Var[S^l]$. Then, from the well-known Chebyshev's inequality, we have the report error requirement as follows for any real number $a > 0$,

$$Pr\{|S^l - E[S^l]| \geq a\} \leq \frac{Var[S^l]}{a^2}. \tag{5.1}$$

Equivalently, Eq. 5.1 can be also expressed as:

$$Pr\{|S^l - E[S^l]| \geq a \times Std[S^l]\} \leq \frac{1}{a^2}, \tag{5.2}$$

where $Std[S^l]$ is the standard deviation of S^l. By the Law of Large Numbers, $E[S^l]$ is close to the expected value or the true value. Then, Eq. 5.2 can be viewed as $Pr\{|error| \geq \sigma_{req}\} \leq p_e$ with $a \geq 1$. As a reminder, σ_{req} is the report error bound and p_e is upper bound on the report error probability.

Both $\sigma_{req} > 0$ and $p_e > 0$ are given as inputs and determined by the application or its operator. By rearranging $p_e = \frac{1}{a^2}$, we have $a = \sqrt{1/p_e} > 0$. Similarly, by rearranging $\sigma_{req} = a \times Std[S^l]$, we can compute the expectation of the least number κ_{req} of sensed readings to be used at the FC when taking a sample mean in order not to violate the report error requirement as follows:

$$E[U^l] \geq \left(\frac{\sqrt{1/p_e} \times \sigma}{\sigma_{req}} \right)^2 = \kappa_{req}. \tag{5.3}$$

In other words, this is the condition guaranteeing that the report error requirement is satisfied on average. In the same manner, κ_{req} is the requirement on the least number of sensed readings to be used to make a sample mean to satisfy the report error requirement on average.

5.4.2 Reporting Interval Requirement

The reporting interval is the time interval for the CTMC to start at state 0 and then return to state 0 for the first time. Please note that there is no direct transition from state 0 to itself. Let τ_{ii} be the smallest return time from state i to state i, i.e., $\tau_{ii} = \inf\{t \geq 0 : X_t = i, X_{t-} \neq i | X_0 = i\}$. Since $\{X_t\}_{t\geq 0}$ is positive recurrent, we have $E[\tau_{ii}] = \frac{1}{\pi_i \times q_i} < \infty$. Thus, τ_{00} is the reporting interval, and from which the reporting interval requirement can be written as: $\tau_{00} \leq \tau_{req}$.

5.4.3 Computing the Optimal Number of Nodes to Activate

For a given μ, as the number of active nodes increases, the expected error of each sample mean decreases since an FC can buffer more data during each reporting interval. Meanwhile, the expected reporting interval decreases because of a shorter holding time in state 0. Let $\eta^*_{ERR}(\mu)$ and $\eta^*_{RI}(\mu)$ be the minimum number of nodes to activate to satisfy the report error and reporting interval (or timely update) requirements, respectively, for a given μ. Then, the minimum number of nodes to activate to satisfy both requirements can be computed as $\eta^*(\mu) = \max\{\eta^*_{ERR}(\mu), \eta^*_{RI}(\mu)\}$.

Furthermore, as μ increases, the expected reporting interval decreases. Thus, $\eta^*_{ERR}(\mu)$ has to increase to guarantee the FC buffers enough number of data within a shorter amount of time. On the other hand, as μ decreases, the expected reporting interval increases. Therefore, $\eta^*_{RI}(\mu)$ has to increase so that the FC can activate more nodes to reduce the holding time in state 0, in order not to violate the maximum reporting interval threshold. To summarize, $\eta^*_{ERR}(\mu)$ is a non-decreasing function of μ, while $\eta^*_{RI}(\mu)$ is non-increasing. Therefore, there exists at least one optimal μ that yields the minimum number of nodes to activate without violating any of the requirements; it will be shown later by Fig. 5.11 in Section 5.6.2. Let μ^* denote such an optimal μ, and let η^* be the smallest number of nodes to activate for μ^*. When there are more than one μ producing the same optimal η^*, we choose the smallest one among the optima as μ^*.

5.5 Node Scheduling Problem and Algorithms

Given the optimal number of nodes to active η^*, we formulate an optimal node activation scheduling problem to maximize the network lifetime by taking the residual energy of nodes into account. In this chapter, the maximum network lifetime is achieved by maximizing the lifetime of each cluster.

5.5.1 Optimal Node Scheduling: Problem Formulation

Nodes are partitioned into a set of groups $\mathcal{G} = \{g_1, g_2, \cdots\}$, and the proposed scheme activates one group at a time. Here, we have $|\mathcal{G}| = G = \lfloor W/\eta^* \rfloor$, where $W = |\mathcal{W}|$. That is, the proposed scheme activates the minimum number (η^*) of nodes to maximize the energy saving. The lifetime of each group is determined by the node with the least amount of residual energy in the group. That is, the lifetime of g_b is proportional to $\min\{\phi(w_a) : \forall w_a \in g_b\}$, where $\phi(w_a)$ is a function that returns the residual energy of node w_a. By maximizing the sum of minimum residual energy over all groups, we can form an optimal node scheduling problem as below (called P. 5.4).

$$\max_{\mathbf{Z}} \sum_{b=1}^{G} \min\{\phi(w_a) : \forall w_a \in g_b\} \tag{5.4a}$$

$$\text{subject to: } \forall b : \sum_{a=1}^{W} z_{ab} = \eta^* \tag{5.4b}$$

$$\forall a : \sum_{b=1}^{G} z_{ab} \leq 1 \tag{5.4c}$$

$$\forall a, b : \; z_{ab} \in \{0, 1\}, \tag{5.4d}$$

where $\mathbf{Z} = [z_{ab}]_{(a,b)\in\{1,2,\cdots,W\}\times\{1,2,\cdots,G\}}$ is a W-by-G membership indicator matrix with $z_{ab} = 1$ if $w_a \in g_b$ or 0 otherwise (i.e., binary relation as in Eq. 5.4d). Each group g_b consists of η^* nodes

(Eq. 5.4b), and each node w_a belongs to at most one group (Eq. 5.4c). We, then, transform P. 5.4 into an MILP formulation as shown below (called P. 5.5).

$$\max_{\mathbf{Z},\overline{\mathbf{Z}},\mathbf{v}} \sum_{b=1}^{G} v_b \tag{5.5a}$$

$$\text{subject to: constraints in } (5.4b), (5.4c), (5.4d)$$

$$\forall a, b: \; v_b \leq \phi(w_a) \times z_{ab} + \phi_{max} \times \overline{z}_{ab} \tag{5.5b}$$

$$\forall a, b: \; z_{ab} + \overline{z}_{ab} = 1, \tag{5.5c}$$

where $\overline{\mathbf{Z}} = \mathbf{1}_{W \times G} - \mathbf{Z}$ (Eq. 5.5c) and ϕ_{max} is the maximum battery capacity of a node. From both Eq. 5.5a and Eq. 5.5b, we have $v_b = \min\{\min\{\phi(w_a) : \forall w_a \in g_b\}, \phi_{max}\}$. The new objective function (Eq. 5.5a) is equivalent to the previous one (Eq. 5.4a) as long as each group has at least one node, which is always guaranteed by Eq. 5.4b. The optimal solutions $\mathbf{Z}^*$ and $\mathbf{v}^* \in \mathbb{R}^G$ indicate which node belongs to which group and the set of the group-wise least residual energy, respectively. The number of binary variables in P. 5.5 is $W \times G \approx W^2/\eta^*$, which could make the problem intractable as the number of nodes increases. However, by carefully studying P. 5.5, we can reduce the number of binary variables.

Let $\Phi = \{\phi_{[1]}, \phi_{[2]}, \phi_{[3]}, \cdots, \phi_{[W]}\}$ be an ordered list of $\phi(w_a)$ for $a = 1, 2, \cdots, W$, such that $\phi_{[k]} < \phi_{[k+1]}$.[2] Also, let $g^{[k]}$ be a group that the minimum residual energy among the nodes in the group is the k^{th} smallest among those values of all groups. For $g^{[k]}$, the set of residual energy of the nodes in the group is denoted by $\Phi^{[k]} = \{\phi^k_{[1]}, \phi^k_{[2]}, \cdots, \phi^k_{[\eta^*]}\}$ such that $\phi^k_{[l]} < \phi^k_{[l+1]}$. For the minimum residual energy, $\phi^k_{[1]}$, of groups $g^{[k]}$ for all $k = 2, 3, \cdots, G-1$, we have $\phi^{k-1}_{[1]} < \phi^k_{[1]} < \phi^{k+1}_{[1]}$.

Claim 1 (Optimal Node Partitioning):[3] Given that Eq. 5.5a equals $\max \sum_{b=1}^{G} \phi^b_{[1]}$, we claim that an optimal partitioning of ϕ values[4] (or an optimal node partitioning) is: $\hat{\Phi}^{[1]} = \{\phi_{[k+1]}, \phi_{[k+2]}, \cdots, \phi_{[k+\eta^*]}\}, \hat{\Phi}^{[2]} = \{\phi_{[k+\eta^*+1]}, \phi_{[k+\eta^*+2]}, \cdots, \phi_{[k+2\eta^*]}\}, \cdots, \hat{\Phi}^{[G]} = \{\phi_{[k+(G-1)\eta^*+1]}, \phi_{[k+(G-1)\eta^*+2]}, \cdots, \phi_{[k+G\eta^*]}\}$, where $k = W - \eta^* \times \lfloor \frac{W}{\eta^*} \rfloor$. The corresponding optimal $\hat{\mathbf{v}}$ vector is:

[2] We assume $\phi_{[k]}$ is real-valued. Thus, without loss of generality, we assume that $\phi_{[k]} \neq \phi_{[k']}$ if $k \neq k'$.

[3] The proof of **Claim 1** is available in Appendix B.

[4] Since each ϕ value is unique, we can derive the partitioning of nodes from the partitioning of ϕ values.

$[\hat{\phi}^1_{[1]}(=\phi_{[k+1]}), \hat{\phi}^2_{[1]}(=\phi_{[k+\eta^*+1]}), \cdots, \hat{\phi}^G_{[1]}(=\phi_{[k+(G-1)\eta^*+1]})]^T$. Also, $\hat{\mathbf{v}}$ is element-wise greater than or equal to any other $\mathbf{v} = \{\phi^1_{[1]}, \phi^2_{[1]}, \cdots, \phi^G_{[1]}\}$, and thus, we have $\sum_{b=1}^{G} \hat{\phi}^b_{[1]} \geq \sum_{b=1}^{G} \phi^b_{[1]}$.

Given the optimal partitioning of nodes, we can make P. 5.5 more tractable by scheduling one group at a time, thus reducing the number of binary variables. The modified problem P. 5.6 chooses η^* nodes to form a group $\hat{g}^{[G]}$, which is the set of nodes with residual energy being $\hat{\Phi}^{[G]}$.

$$\max_{\mathbf{z},\overline{\mathbf{z}},v} \; v \tag{5.6a}$$

$$\text{subject to:} \; \sum_{a=1}^{W} z_a = \eta^* \tag{5.6b}$$

$$\forall a: \; z_a \in \{0,1\} \tag{5.6c}$$

$$\forall a: \; v \leq \phi(w_a) \times z_a + \phi_{max} \times \overline{z_a}, \tag{5.6d}$$

$$\forall a: \; z_a + \overline{z}_a = 1, \tag{5.6e}$$

where $\mathbf{z} \in \{0,1\}^W$ denotes the membership relation between each node and $\hat{g}^{[G]}$, $\overline{\mathbf{z}} = \mathbf{1}_W - \mathbf{z}$, and $v \in \mathbb{R}$ is the minimum nodal residual energy in $\hat{g}^{[G]}$, i.e., $\hat{\phi}^G_{[1]}$ or $\phi_{[k+(G-1)\eta^*+1]}$. In P. 5.6, the number of binary variables is W, which is much smaller than W^2/η^* for a large W and a small η^*.

To summarize, we have formulated an optimal node scheduling problem P. 5.4 which partitions all nodes in a cluster into G groups. The problem is then converted to an MILP formulation which is P. 5.5. Finally, based on the analytical solution to P. 5.5 given by **Claim 1**, we have formulated an optimal node scheduling problem P. 5.6 which chooses only η^* number of nodes to activate to reduce the complexity of the combinatorial optimization problem.

5.5.2 Optimal and Greedy Node Scheduling Algorithms

Given the optimal node scheduling for $\hat{g}^{[G]}$ from P. 5.6, we can derive an Optimal Node Scheduling (ONS) algorithm which runs as follows. [Step 1] ONS runs P. 5.6 to choose η^* number of nodes to activate. [Step 2] When any of the active nodes does not have enough residual energy to operate, remove/deactivate the node, and go back to [Step 1]. Please note that P. 5.6 is still an MILP. When there are a small number of binary variables (e.g., < 50), an optimal solution can be found in a

Algorithm 6 Optimal/Greedy Node Scheduling

```
1: if $|\mathcal{W}| < \eta^*$ then
2:     Terminate                                   // infeasible, not enough nodes
3: end if
4: $g \leftarrow \{\}$                              // initialize to an empty set
5: if ONS is chosen then
6:     $\mathbf{z}^* \leftarrow$ Solve P. 5.6
7:     $g \leftarrow g \cup \{w_a\}, \forall a$ such that $z_a^* = 1$
8: else if GNS is chosen then
9:     while $|g| < \eta^*$ do
10:        $a^* \leftarrow \arg_a \max\{\phi(w_a) : \forall a \in \texttt{idx}(\mathcal{W} - g)\}$
11:        $g \leftarrow g \cup \{w_{a^*}\}$
12:    end while
13: end if
14: while $|g| = \eta^*$ do
15:    Activate nodes in $g$
16:    if $\exists w_a \in g$ such that $\phi(w_a) < \phi_{thr}$ then
17:        $g \leftarrow g - \{w_a\}$, $\mathcal{W} \leftarrow \mathcal{W} - \{w_a\}$
18:    end if
19: end while
20: Go to line:1
```

negligible amount of time, but it is not always the case if the number of nodes in a cluster is large or the processing power of the FC is limited.

To further reduce the complexity, we propose a Greedy Node Scheduling (GNS) algorithm that produces the same node selection as ONS but with a much lower complexity of O($\eta^* W$). Given that an optimal node selection for $\hat{g}^{[G]}$ is equivalent to finding nodes with residual energy of $\hat{\Phi}^{[G]}$ = $\{\phi_{[k+(G-1)\eta^*+1]}, \phi_{[k+(G-1)\eta^*+2]}, \cdots, \phi_{[k+G\eta^*]}\}$, this can be done by choosing η^* nodes with the highest residual energy. Please note that $\phi_{[k+(G-1)\eta^*+1]}, \phi_{[k+(G-1)\eta^*+2]}, \cdots, \phi_{[k+G\eta^*]}$ correspond to $(\eta^*)^{\text{th}}$, $(\eta^*-1)^{\text{st}}$, $\cdots$, 1^{st} highest nodal residual energy, respectively, in a cluster. GNS has the same two-step algorithm as ONS except in [Step 1], instead of solving P. 5.6, GNS iteratively selects a node with the n^{th} highest nodal residual energy where $n = 1, 2, \cdots, \eta^*$ to activate. The algorithm for ONS and GNS is given in Algo. 6.

Algo. 6 first checks if the number of nodes in the cluster is at least η^* (line 1); if not, it terminates (line 2). Then, the set g is initialized to an empty set (line 4). For ONS, g is determined by the solution to P. 5.6 (lines 5–7). On the other hand, GNS chooses nodes to activate by iteratively searching for a node with the highest residual energy (lines 8–12). At line 10, the $\texttt{idx}(\cdot)$ function returns the index set; e.g., $\texttt{idx}(\{w_1, w_2, w_7\}) = \{1, 2, 7\}$. When g is determined, all nodes in the group become activated (line 15), while the others stay in the sleep mode to minimize the energy consumption. The ϕ_{thr} (lines 16) is the battery threshold such that if the residual energy of a node drops below the threshold, the node can no longer function properly. Any battery-drained nodes will be removed from both g and $\mathcal{W}$ (line 17), and the algorithm starts over (line 20) from the beginning.

5.6 Performance Evaluation

We have implemented the following five different node activation scheduling algorithms and compared their performances with each other: ONS (Optimal Node Scheduling), GNS (Greedy Node Scheduling), RNS (Random Node Scheduling), SNS (Sequential Node Scheduling), and invGNS (inverse GNS). RNS activates nodes uniformly at random, SNS activates nodes with the smallest AID[5] first (which is similar to *first-come, first-served*), and invGNS activates nodes with the least amount of nodal residual energy first, which is the exact opposite of GNS. Please note that all five algorithms activate the same η^* number of nodes at a time for a fair comparison. We used Matlab [49] to solve the optimization problem P. 5.6 as well as to implement a simulator and the five node scheduling algorithms.

5.6.1 Network Configuration

Unless stated otherwise, we use the following default simulation configuration. The FC is associated with $|\mathcal{W}| = 50$ nodes, and it can buffer up to $B = 50$ data. Each node has a fixed sensing and transmit rate of $0.3 = \lambda/|\mathcal{W}|$ (data per second, from a node to its associated FC). The

[5] As a reminder, when a node joins a 802.11ah-compatible network, it is assigned a unique identifier, called AID.

variance of the sensed readings (i.e., the measurement error) is $\sigma^2 = 25$. We assume the following requirements:

- $\sigma_{req} = \sigma$,
- $p_e = 0.1$ (i.e., 10%), and
- $\tau_{req} = 6$ seconds (i.e., the AP expects at least one report every six seconds on average).

From the first two requirements, we can also compute the expected number of sensed readings to be buffered at the FC and to be used to take a sample mean in order not to violate the report error requirement. After plugging in p_e, σ and σ_{req} into Eq. 5.3, we get $E[U^l] = 10$, meaning that, on each interval, the FC needs to take a sample mean out of 10 or more data on average for the report error requirement. Therefore, from now on we use $\kappa_{req} = 10$.

For μ (i.e., the rate at which the FC shall transmit a report to the AP), we have $\mu_{\min} = 0.1$, $\mu_{\max} = 0.5$, and $\Delta_\mu = 0.001$. Considering that the radio module is one of the major sources of energy consumption [118], we assume the following simplified energy model for nodes. At the beginning, each node has a random amount of energy drawn from a uniform distribution, $U(0, \phi_{\max}]$, where $\phi_{\max} = 100$. Each node consumes one unit of power for generating one sensed reading or transmitting one sensed reading to the FC; in this regard, ϕ_{thr} is set to 1. All results are averaged over 50 simulation runs. We also have measured the 95% confidence intervals of the results, which, however, are omitted from the figures as they are very small.

5.6.2 Computing the Optimal μ^* and η^*

We first show how to compute the optimal μ^* and η^*. The analytical results in Fig. 5.8 and Fig. 5.9 show the expected duration of the reporting interval and the number of data to be used to take a sample for each reporting interval, respectively, as the number of active nodes increases. As shown in Fig. 5.8, with an increased number of active nodes, the expected duration of a reporting interval decreases for each given μ. This is because the FC spends less and less time in state 0 as the total incoming data rate increases. In fact, it approaches $1/\mu$ as the number of active nodes

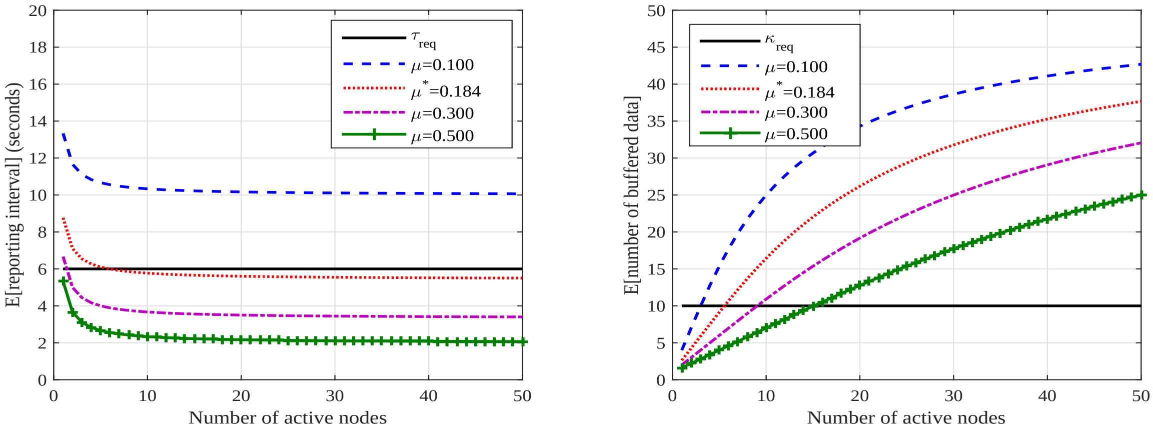

Figure 5.8 The expectation of the reporting interval decreases as the number of active nodes increases for each given μ.

Figure 5.9 The expectation of the number of data to be used by the FC to take a sample mean increases with the number of active nodes for each given μ.

Figure 5.10: Expectation of the length of the reporting interval and the number of data buffered in a reporting interval with respect to the number of active nodes.

increases. The reporting interval τ_{00} is the sum of the holding time in state 0 (denoted by τ_0) and the sum of holding times in non-zero states (denoted by τ_Σ) until the CTMC returns to the state 0 for the first time. While τ_Σ is a function of both μ and λ_{act}, where λ_{act} = (the number of active nodes) × $\lambda/|\mathcal{W}|$, τ_0 is dependent only on λ_{act} since there is no transition from state 0 to itself. As the number of active nodes increases, so does the rate of data arrivals from nodes to the FC. Thus, the waiting time until the first data arrival also decreases, resulting in the reduced reporting interval. As shown in Fig. 5.8, if μ is too small, e.g., 0.100, the reporting interval requirement cannot be satisfied at all. On the other hand, for larger μ's, we can identify the smallest number of active nodes, on and beyond which the reporting interval requirement is satisfied.

On the other hand, as shown in Fig. 5.9, the expectation of the number κ_{req} of data to be buffered and to be used to take a sample mean at the FC increases with the number of active nodes for each given μ. This is because having more active nodes increases the data arrival rate to the FC. In contrast to the reporting interval, a cluster was able to satisfy the κ_{req} requirement for any μ available, i.e., from 0.100 to 0.500. By the Law of Large Numbers, taking more data to compute the sample mean produces a more accurate report. Therefore, as the number of active nodes increases the error in each report (or sample mean) decreases.

Since both the reporting interval and report error requirements have to be satisfied at the same time for a given μ, we first take the smallest number of active nodes needed for each requirement separately, and then, take the larger one to find $\eta^*(\mu)$ for a given μ, i.e., $\eta^*(\mu) = \max\{\eta^*_{ERR}(\mu), \eta^*_{RI}(\mu)\}$. The trace of $\eta^*(\mu)$ for all available $\mu \in [\mu_{\min}, \mu_{\max}]$ is shown in Fig. 5.11. For small values of $\mu \in [\mu_{\min}, 0.168]$, the FC cannot satisfy the reporting interval requirement, and thus, η^* is set to 0 to signify infeasibility. When $\mu > 0.168$, there exists a feasible $\eta^*(\mu)$ for each μ forming a convex-like shape, and in the end, we have found that $\mu^* = 0.184$ to yield the smallest $\eta^* = 6$. Here, any μ yielding $\eta^* = 6$ can be an optimal solution; yet, in the proposed scheme, the minimum μ among optima is chosen.

5.6.3 Performance Comparison of the Node Scheduling Algorithms

We also have compared the cluster lifetime with respect to the number of nodes in a cluster, and the results are plotted in Fig. 5.13. As shown in the figure, both ONS and GNS result in the longest lifetime than other schemes. Also, the performance of GNS equals that of ONS with a much lower complexity. The invGNS has the worst performance as it activates node in the exactly opposite way of the GNS. The RNS outperforms SNS, showing that the randomization in node activation yields a better performance than both SNS and invGNS. In addition, we have compared the cluster lifetime with different maximum battery capacity when $|\mathcal{W}| = 50$. That is, the initial nodal energy of each node is drawn uniformly at random from the range $(0, \phi_{\max}]$, where $\phi_{\max} = 100, 300, \cdots, 900$. The results are shown in Fig. 5.14. Again, both ONS and GNS outperform the rest with invGNS being the worst.

5.7 Conclusion

In this chapter, we studied the problem of activation scheduling of IoT devices for environment monitoring applications. We first analyzed the network of our interest by using a Markov process.

From the stationary distribution of CTMC, we computed the minimum number of devices to activate to satisfy both report error and timely update requirements. Then, we derived an optimal scheduling algorithm for device activation that maximizes the cluster lifetime by taking into consideration the nodal residual energy. We also proposed a greedy scheduling algorithm to produce the optimal performance but with a significantly lower complexity. Evaluation results show that the proposed schemes, ONS and GNS, can effectively prolong the network lifetime.

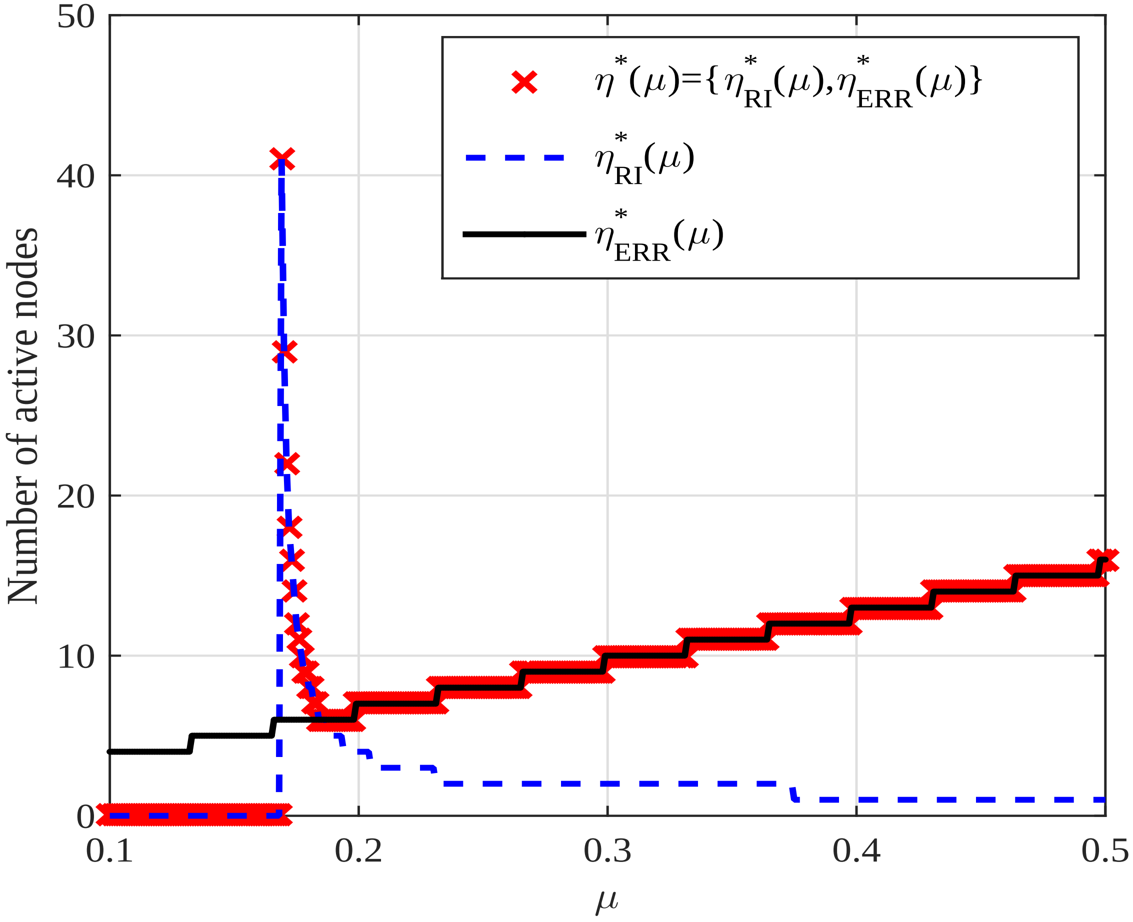

Figure 5.11: Optimal number of nodes $\eta^*(\mu)$ to activate with respect to μ, where $\eta^*_{RI}(\mu)$ and $\eta^*_{ERR}(\mu)$ are the smallest number of nodes to activate to satisfy reporting interval and report error requirements, respectively, and $\mu^* = 0.184$.

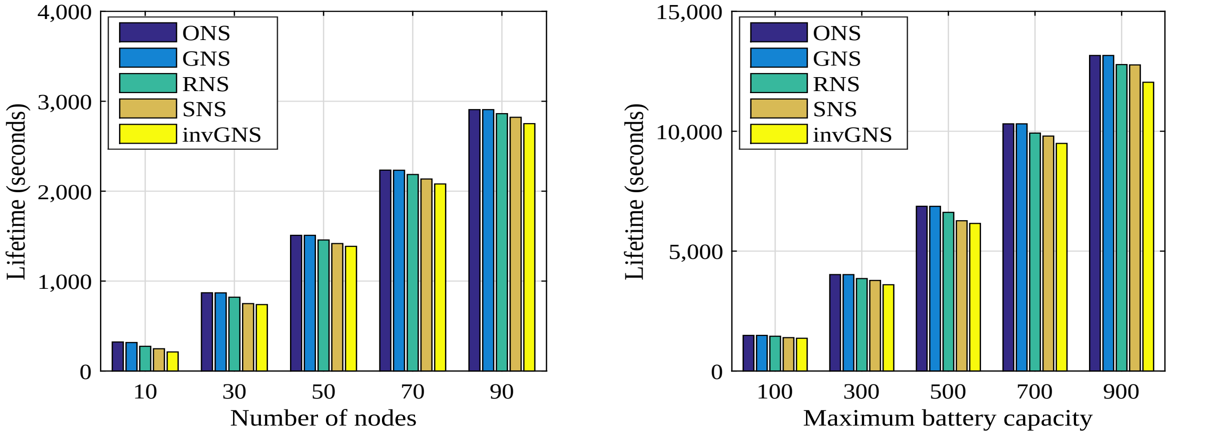

Figure 5.13 With respect to the number of nodes in a cluster.

Figure 5.14 With respect to the maximum battery capacity of nodes.

Figure 5.15: Comparison of the cluster lifetime with five different node scheduling methods.

CHAPTER 6. CONCLUDING REMARKS

In conclusion, this chapter summarizes the studies introduced in this dissertation from Chapter 2 to Chapter 5, and then, provides an overview of the future research directions.

6.1 Summary

In this dissertation, we have studied two networking systems, and for each we proposed an optimal resource scheduling scheme from the perspective of energy efficiency. The discussed topics are expected to play an important role in enhancing the current networking technologies as well as making progress towards the next generation networks.

In Chapter 2, we have proposed a framework for a heterogeneous cellular network with which a network operator can minimize the energy consumption while satisfying users' quality of service demand. By utilizing low-power smallcell base stations, users' traffic demand can be satisfied with less amount of energy consumption compared to the conventional cellular networks consisting only of macro base stations. Given that the joint user association and resource allocation problem is nonconvex, the problem formulation is split into two in order for it to be tractable. As a result, we have an association problem and a resource scheduling problem. A relaxation technique has been used to further reduce the problem complexity, and then, we have proposed an iterative method to find an optimal resource allocation strategy. The proposed method is lightweight and can be operated in a distributed manner, which all together allows the proposed method to operate in realtime.

In Chapter 3, we have studied the resource optimization problem for C-RAN. By leveraging the powerful computing capability and the centrality, a C-RAN can effectively allocate and utilize the computation resource as well as the rest of the networking resources in a cost-efficient manner. In addition, by taking advantage of the large coverage of macro base stations, heterogeneous C-RAN

(H-CRAN) can further increase the resource utilization. In H-CRANs, RRHs do not need to be densely deployed, alleviating the severe inter-tier interference problem between RRHs. In order to effectively consider the uncertainties in both users' mobility and their varying service demand, we have proposed a multi-stage stochastic programming based resource provisioning and service strategy. We have also proposed a new operation mode in which mobile network operators can share their infrastructure in order to increase the overall resource utilization, while minimizing the service outage. The proposed method optimizes a series of resources, including the VBS allocation for each RRH, user-RRH association, channel allocation for users and resource allocation from a different mobile network operator and its optimal usage, so that the mobile network operator can maximize its profit while minimizing energy consumption.

In Chapter 4, we have studied a network with IEEE 802.11ah which is considered as a promising WLAN standard for connecting a large number of IoT devices in a sub-1 GHz unlicensed band. Despite the fact that an 802.11ah access point can associate with and manage a large number of the resource-constrained, battery-operated devices in an energy-efficient and structured manner, we have identified an energy waste problem that may reduce the lifetime the IoT devices and the entire network. In order to minimize the energy waste from IoT devices without degrading the overall network throughput performance, we have proposed an enhanced energy saving mechanism for 802.11ah. Temporary membership change, which is the main idea of the proposed method, allows IoT devices to switch between groups, and with which an 802.11ah network can increase the number of devices that can stay in the low power state without being interrupted. The proposed method is not only energy-efficient, but also tightly tailored for 802.11ah. The proposed traffic re-ordering algorithm is low complex, and it does not increase the average delay in data delivery.

In Chapter 5, we have studied the effect of the sensing schedule (or the device activation schedule) of battery-operated devices on the energy efficiency. We also have considered an IEEE 802.11ah network with an environment monitoring application. By analyzing the network we have derived the optimal configuration for the network, i.e., the optimal reporting rate of the fusion center and the optimal number of devices to activate to satisfy the application-specific requirements.

Also, we have formulated a residual-energy-aware optimal device activation scheduling problem to maximize the network lifetime. To reduce the time-complexity of the optimal scheduling scheme, we proposed a low-complex heuristic algorithm of which performance is comparable to the optimum. The evaluation results confirmed that the proposed optimal scheduling method and the equivalent heuristic algorithm outperform the conventional device activation algorithms and maximize the network lifetime without violating any of the system requirements.

6.2 Future Research Directions

The fundamental problems and challenges addressed in this dissertation build a foundation on the research for the next-generation networking systems and applications. The cloud-based networking system will play a key role in the future networks for its high processing power. Due to the enormous volume of data to be collected from a large-scale IoT sensor network and mobile users, the machine learning and artificial techniques need to be further developed to automate the procedures such as network resource scheduling. Such a transition to the automated machine learning based decision making process may provide a better understanding on the interaction between the network and its serving users, revealing new applications and opportunities.

The cloud-based networking architecture C-RAN can provide an energy-efficient cellular service as discussed in Chapter 3. As we have seen in both Chapter 4 and Chapter 5, the energy-efficient operation of IoT devices can prolong the network lifetime, and as a result, data can be accumulated for longer period of time. The next goal is to combine the two networks together so that we can expedite the era of the fully-connected future networks. In this regard, we plan to continue carrying out research in the following three interesting topics.

6.2.1 Unmanned Aerial Vehicle (UAV)-Assisted Network for Better Connectivity

IoT has a variety of applications, and one of which is to monitor a remotely-located and possibly hazardous area. For such networks, it may not be a feasible plan to provide an additional network infrastructure to connect the IoT network and the central computing and/or data center. One possible solution is to utilize the UAVs (or drones) with a long-range cellular communication interface to connect the two separated networks by relaying the sensed readings collected from IoT networks. Optimizing the multi-hop relaying transmissions over a mesh UAV network is challenging for their high mobility. What makes the problem more complicated is the delay constraints in delivering the sensed readings. In some applications such as disaster monitoring, timely update of the changes in the environment is very important. In this regard, we plan to propose an optimal data relay and an interference management scheme for UAV networks, given the time constraints in data delivery.

6.2.2 Self-Configuration and Self-Healing Networks

As the network evolves into a much more complex structure (e.g., C-RAN), the traditional network modeling techniques may not be a feasible solution to resource optimization. Severe simplification of the system may allow the conventional resource optimization approach to be applicable, but the solution driven from such simplified models will be impractical and inefficient. One effective solution is to let network optimize its resource usage by itself. Due to the larger set of data collected from the network infrastructure and IoT devices as well as the powerful computing resources available, a machine-learning based self-configuration approach can become more effective in resource optimization. In addition, such an automated learning can be used to detect the failure of the network infrastructure (e.g., base station) with a short delay. In this regard, we propose to design a self-configuring and self-healing mechanism for the future network system to provide a resource-efficient and breakout-free networking service to users.

6.2.3 Mobile Edge Computing for Tactile Internet with Ultra-Low Delay

The Tactile Internet (TI) is a promising application, and is considered as the next-generation IoT. However, TI can be realizable only by combining the IoT network and H-CRANs (or C-RANs). One major issue in realizing the TI is to guarantee a ultra-low end-to-end delay. In some applications, the Tactile Internet requires a round-trip delay of 1 ms [7]. Although H-CRANs can provide a high availability with reliable communication links, it may not be able to fulfill the delay requirement. For example, the fronthaul link between RRHs and the BBU pool can be a major source of causing a delay. One practical solution is to put the processing units close to the end-users, which is called Mobile Edge Computing (MEC) [128]. Some of the processing units can be placed at the front-end base stations (i.e., RRHs), and then, the delay-sensitive user requests (or tasks) can be processed therein. Such an architecture can significantly reduce the end-to-end delay. Also, it can reduce the amount of traffic injected to the network, alleviating the network congestion. In this regard, we plan to propose an optimal algorithm to forward the user requests to either the central processing units or MEC units in order to satisfy the end-to-end delay constraints.

APPENDIX A. THREE-POINT APPROXIMATION

In order to discretize a uniform random variable into three discrete points, we have used the method proposed in [79]. Let G be the distribution of continuous Uniform(a,b) with mean m. Also, let G^d be the distribution of the discretized G. By using the mass transportation problem framework [79], the objective of the discretization is to minimize the distance d between the two distributions which is defined as:

$$d(G, G^d) = \sum_{i=1}^{K} \int_{\frac{z_{i-1}-z_i}{2}}^{\frac{z_i-z_{i+1}}{2}} |u - z_i| \; dG(u),$$

where the number of points K to discretize the original distribution into is 3. That is, $G^d = \{z_1, z_2 = m, z_3\}$ and $z_i \in \mathbb{R}$ for $\forall i$. Also, z_0 and z_4 are $-\infty$ and $+\infty$, respectively. Since the service demand cannot be negative, we have $a = 0$.

Then, the discretization (or 3-point approximation) problem can be written as: $\min_{\mathbf{z}}. \; d(G, G^d) = \int_{\frac{-\infty+z_1}{2}}^{\frac{z_1+z_2}{2}} |u - z_1| \; dG(u) + \int_{\frac{z_1+z_2}{2}}^{\frac{z_2+z_3}{2}} |u - z_2| \; dG(u) + \int_{\frac{z_2+z_3}{2}}^{\frac{z_3+\infty}{2}} |u - z_3| \; dG(u)$. After manipulating the right-hand side, we can transform it into a simple form: $\mathbf{z}^T\mathbf{Q}\mathbf{z} + f(\mathbf{z})$, where $\mathbf{z} = [z_1, z_2, z_3]^T$, $f(\mathbf{z})$ is an affine function of $\mathbf{z}$, and $\mathbf{Q}$ is symmetric and positive definite. Thus, there exists a unique optimal solution to the following quadratic problem (called P. A.1).

$$\min_{\mathbf{z}}. \; \mathbf{z}^T\mathbf{Q}\mathbf{z} + f(\mathbf{z}) \tag{A.1a}$$

subject to

$$\mathbf{z} \succeq \mathbf{0}, \tag{A.1b}$$

$$a \le z_1 \le z_2 \le z_3 \le b, \tag{A.1c}$$

$$z_2 = m, \tag{A.1d}$$

where $\succeq$ is an element-wise greater than or equal to operator. The solution is non-negative (Eq. A.1b), bounded by the lower and upper bound, i.e., a and b, respectively (Eq. A.1c), and finally, the mean value should remain the same (Eq. A.1b) as that of the original distribution. The optimal solution, **z**, is a vector of three equally likely discrete values that are approximations of the original uniform distribution with the same mean value.

APPENDIX B. PROOF OF CLAIM 1

Let us assume W is a multiple of η^*. Then, the optimal partitioning of ϕ values and the corresponding group-wise least residual energy vector are $\hat{\Phi}^{[1]}, \hat{\Phi}^{[2]}, \cdots, \hat{\Phi}^{[G]}$ and $\hat{\mathbf{v}}$, respectively, as given in Chapter 5 **Claim 1** with $k = W - \eta^* \times \lfloor \frac{W}{\eta^*} \rfloor = 0$. That is, we have $\hat{\mathbf{v}} = [\hat{\phi}^1_{[1]}(= \phi_{[1]}), \hat{\phi}^2_{[1]}(= \phi_{[\eta^*+1]}), \cdots, \hat{\phi}^G_{[1]}(= \phi_{[(G-1)\eta^*+1]})]^T$.

Proof by contradiction:

Suppose there exists a partitioning of ϕ's, $\tilde{\Phi}^{[1]}, \tilde{\Phi}^{[2]}, \cdots, \tilde{\Phi}^{[G]}$, with the corresponding group-wise least residual energy vector $\tilde{\mathbf{v}} = [\tilde{\phi}^1_{[1]}, \tilde{\phi}^2_{[1]}, \cdots, \tilde{\phi}^G_{[1]}]^T$ such that $\sum_{b=1}^{G} \hat{\phi}^b_{[1]} < \sum_{b=1}^{G} \tilde{\phi}^b_{[1]}$. Then, there exists at least one b' such that $\hat{\phi}^{b'}_{[1]}(= \phi_{[(b'-1)\eta^*+1]}) < \tilde{\phi}^{b'}_{[1]}$. Let b' be the smallest group index among the groups satisfying the inequality, i.e., $b' = \min \arg_{b=1,2,\cdots,G}\{\hat{\phi}^b_{[1]} < \tilde{\phi}^b_{[1]}\}$. Then, it is always the case that $\hat{\phi}^1_{[1]}(= \phi_{[1]}) = \tilde{\phi}^1_{[1]}$ since $\phi_{[1]}$ is the smallest ϕ, and thus, b' cannot be 1. Also, we have $\hat{\phi}^G_{[1]}(= \phi_{[(G-1)\eta^*+1]}) \geq \tilde{\phi}^G_{[1]}$ since $\phi_{[(G-1)\eta^*+1]}$ is the largest ϕ among those that can be chosen as the minimum residual battery in a group. Therefore, b' cannot be G. Now, let b' be any integer in the range $(1, G)$. Since all ϕ values in $\tilde{\Phi}^{[b']}, \tilde{\Phi}^{[b'+1]}, \cdots, \tilde{\Phi}^{[G]}$ are greater than or equal to $\tilde{\phi}^{b'}_{[1]}$, $\phi_{[(b'-1)\eta^*+1]}(= \hat{\phi}^{b'}_{[1]})$ has to be a member of one of $\tilde{\Phi}^{[1]}, \tilde{\Phi}^{[2]}, \cdots, \tilde{\Phi}^{[b'-1]}$. That means at least one ϕ satisfying $\phi_1 < \phi < \phi_{[(b'-1)\eta^*+1]}$ belongs to one of $\tilde{\Phi}^{[b']}, \tilde{\Phi}^{[b'+1]}, \cdots, \tilde{\Phi}^{[G]}$, which contradicts that all ϕ's in $\tilde{\Phi}^{[b']}, \tilde{\Phi}^{[b'+1]}, \cdots, \tilde{\Phi}^{[G]}$ are greater than or equal to $\tilde{\phi}^{b'}_{[1]}$. In general, when W is not a multiple of η^*, the optimal partitioning of ϕ values and the corresponding optimal vector of per-group minimum residual batteries are $\hat{\Phi}^{[1]}, \hat{\Phi}^{[2]}, \cdots, \hat{\Phi}^{[G]}$ and $\hat{\mathbf{v}}$ as given in Chapter 5 **Claim 1** with $k = W - \eta^* \times \lfloor \frac{W}{\eta^*} \rfloor > 0$. Since $\phi_{[1]}, \phi_{[2]}, \cdots, \phi_{[k]}$ are smaller than the rest, if any of these belong to a group (or groups), they reduce the minimum residual energy of the group (or groups) they belong to compared to $\hat{\mathbf{v}}$, and thus, it yields a smaller sum of the group-wise least residual energy values than that of $\hat{\mathbf{v}}$. ■

BIBLIOGRAPHY

[1] J. O'Toole, "Mobile apps overtake PC Internet usage in U.S.," *CNN Money*, 2014 [Online]. Available: http://money.cnn.com/2014/02/28/technology/mobile/mobile-apps-internet/. [Accessed: Apr. 10, 2017]

[2] T. Bindi, "Mobile and Tablets Internet Usage Surpasses Desktop for First Time: StatCounter," *ZDNet*, 2016 [Online]. Available: http://www.zdnet.com/article/mobile-and-tablet-internet-usage-exceeds-desktop-for-first-time-statcounter/. [Accessed: Apr. 10, 2017]

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

[4] "Cisco visual networking index: Global mobile data traffic forecast update, 2014–2019," Cisco Systems, Inc., San Jose, CA, Cisco White Paper, Feb. 2015.

[5] "5G systems," Ericsson, Stockholm, Sweden, Ericsson White Paper, Jan. 2017.

[6] L. Atzori, A. Iera, and G. Morabito, "From Smart Objects to Social Objects: The Next Evolutionary Step of the Internet of Things," *IEEE Communications Magazine*, vol. 52, no. 1, pp. 97–105, Jan. 2014.

[7] A. Aijaz, M. Dohler, A. H. Aghvami, V. Friderikos, and M. Frodigh, "Realizing the Tactile Internet: haptic communications over next generation 5G cellular networks," *IEEE Wireless Communications*, vol. 24, no. 2, pp. 82–89, Apr. 2017.

[8] "TGah Functional Requirements and Evaluation Methodology Rev.5," IEEE 802.11-09/00000905r5, Jan. 2012.

[9] "IEEE P802.11 Wireless LANs: Specification Framework for TGah," IEEE 802.11-11/1137r15, May 2013.

[10] R. Bolla, R. Bruschi, F. Davoli, and F. Cucchietti, "Energy efficiency in the future Internet: A survey of existing approaches and trends in energy-aware fixed network infrastructures," *IEEE Communications Surveys & Tutorials*, vol. 13, no. 2, pp. 223–244, July 2011.

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

[12] M. Peng, Y. Sun, X. Li, Z. Mao, and C. Wang, "Recent Advances in Cloud Radio Access Networks: System Architectures, Key Techniques, and Open Issues," *IEEE Communications Survey & Tutorials*, vol. 18, no. 3, pp. 2282–2308, Mar. 2016.

[13] T. Kim and J. M. Chang, "QoS-Aware Energy Efficient Association and Resource Scheduling for HetNets," *IEEE Transactions on Vehicular Technology*, vol. 67, no. 1, pp. 650–664, Jan. 2018.

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

[15] P. Demestichas, A. Georgakopoulos, D. Karvounas, K. Tsagkaris, V. Stavroulaki, J. Lu, C. Xiong, and J. Yao, "5G on the horizon: Key challenges for the radio-access network," *IEEE Vehecular Technology Magazine*, vol. 8, no. 3, pp. 47–53, July 2013.

[16] W. H. Chin, Z. Fan, and R. Haines, “Emerging technologies and research challenges for 5G wireless networks,” *IEEE Wireless Communications*, vol. 21, no. 2, pp. 106–112, Apr. 2014.

[17] T. Q. S. Quek, G. de la Roche, I. Guvenc, and M. Kountouris, *Small cell networks: deployment, PHY techniques, and resource management*, Cambridge, United Kingdom: Cambridge University Press, 2013.

[18] 3rd Generation Partnership Project (3GPP), Evolved universal terrestrial radio access (E-UTRA) and evolved universal universal terrestrial radio access network (E-UTRAN); Overall description; Stage 2, TS 36.300, Release 13, June 2016.

[19] D. Astely, E. Dahlman, G. Fodor, S. Parkvall, and J. Sachs, “LTE release 12 and beyond [Accepted From Open Call],” *IEEE Communications Magazine*, vol. 51, no. 7, pp. 154–160, July 2013.

[20] I. Ashraf, F. Boccardi, and L. Ho, “SLEEP mode techniques for small cell deployments,” *IEEE Communications Magazine*, vol. 49, no. 8, pp. 72–79, Aug. 2011.

[21] Y. Li, M. Sheng, C. W. Tan, Y. Zhang, Y. Sun, X. Wang, Y. Shi, and J. Li, “Energy-efficient subcarrier assignment and power allocation in OFDMA systems with max-min fairness guarantees,” *IEEE Transactions on Communications*, vol. 63, no. 9, pp. 3183–3195, Sep. 2015.

[22] L. Venturino, A. Zappone, C. Risi, and S. Buzzi, “Energy-efficient scheduling and power allocation in downlink OFDMA networks with base station coordination,” *IEEE Transactions on Wireless Communications*, vol. 14, no. 1, pp. 1–14, Jan. 2015.

[23] Y.-P. Zhang, S. Feng, and P. Zhang, “Adaptive cell association and interference management in LTE-A small-cell networks,” *IEEE Vehicular Technology Conference (VTC Fall)*, Las Vegas, NV, 2013, pp. 1–6.

[24] E. Aryafar, A. Keshavarz-Haddad, M. Wang, and M. Chiang, "RAT selection games in HetNets," *IEEE International Conference on Computer and Communications (INFOCOM)*, Turin, Italy, 2013, pp. 998–1006.

[25] S. Singh and J. G. Andrews, "Joint resource partitioning and offloading in heterogeneous cellular networks," *IEEE Transactions on Wireless Communications*, vol. 13, no. 2, pp. 888–901, Dec. 2014.

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

[27] W. Wang, X. Wu, L. Xie, and S. Lu, "Femto-matching: efficient traffic offloading in heterogeneous cellular networks," *IEEE International Conference on Computer and Communications (INFOCOM)*, Hong Kong, China, 2015, pp. 325–333.

[28] K. Shen and W. Yu, "Distributed pricing-based user association for downlink heterogeneous cellular networks," *IEEE Journal on Selected Areas in Communications*, vol. 32, no. 6, pp. 1100–1113, June 2014.

[29] V. N. Ha and L. B. Le, "Fair resource allocation for OFDMA femtocell networks with macrocell protection," *IEEE Transactions on Vehicular Technology*, vol. 63, no. 3, pp. 1388–1401, Oct. 2014.

[30] D. T. Ngo, S. Khakurel, and T. Le-Ngoc, "Joint subchannel assignment and power allocation for OFDMA femtocell networks," *IEEE Transactions on Wireless Communications*, vol. 13, no. 1, pp. 342–355, Dec. 2014.

[31] W. C. Cheung, T. Q. S. Quek, and M. Kountouris, “Throughput optimization, spectrum allocation, and access control in two-tier femtocell networks,” *IEEE Journal on Selected Area in Communications*, vol. 30, no. 3, pp. 561–574, Apr. 2012.

[32] W. Bao and B. Liang, “Radio resource allocation in heterogeneous wireless networks: A spatial-temporal perspective,” *IEEE International Conference on Computer and Communications (INFOCOM)*, Hong Kong, China, 2015, pp. 334–342.

[33] D. Fooladivanda and C. Rosenberg, “Joint resource allocation and user association for heterogeneous wireless cellular networks,” *IEEE Transactions on Wireless Communications*, vol. 12, no. 1, pp. 248–257, Dec. 2013.

[34] V. Chandrasekhar and J. G. Andrews, “Spectrum allocation in tiered cellular networks,” *IEEE Transactions on Communications*, vol. 57, no. 10, pp. 3059–3068, Oct. 2009.

[35] B. Zhuang, D. Guo, and M. L. Honig, “Traffic-driven spectrum allocation in heterogeneous networks,” *IEEE Journal on Selected Areas in Communications*, vol. 33, no. 10, pp. 2027–2038, May 2015.

[36] A. Abdelnasser, E. Hossain, and D. I. Kim, “Tier-aware resource allocation in OFDMA macrocell-small cell networks,” *IEEE Transactions on Communications*, vol. 63, no. 3, pp. 695–710, Feb. 2015.

[37] Y. Li, T. Jiang, M. Sheng, and Y. Zhu, “QoS-aware admission control and resource allocation in underlay device-to-device spectrum-sharing networks,” *IEEE Journal on Selected Areas in Communications*, vol. 34, no. 11, pp. 2874–2886, Nov. 2016.

[38] K. Son, S. Lee, Y. Yi, and Song Chong, “REFIM: A practical interference management in heterogeneous wireless access networks,” *IEEE Journal on Selected Areas in Communications,* vol. 29, no. 6, pp. 1260–1272, Jun. 2011.

[39] Y. Li, M. Sheng, Y. Sun, and Y. Shi, "Joint Optimization of BS Operation, User Association, Subcarrier Assignment, and Power Allocation for Energy-Efficient HetNets," *IEEE Journal on Selected Areas in Communications*, vol. 34, no. 12, pp. 3339–3353, Dec. 2016.

[40] G. Bacci, E. V. Belmega, P. Mertikopoulos, and L. Sanguinetti, "Energy-aware competitive power allocation for heterogeneous networks under QoS constraints," *IEEE Transactions on Wireless Communications*, vol. 14, no. 9, pp. 4728–4742, Apr. 2015.

[41] W. Saad, Z. Han, R. Zheng, M. Debbah, and H. V. Poor, "A college admission game for uplink user association in wireless small cell networks," *IEEE International Conference on Computer and Communications (INFOCOM)*, Toronto, ON, 2014, pp. 1096–1104.

[42] 3rd Generation Partnership Project (3GPP), Evolved universal terrestrial radio access (E-UTRA); further advancements for E-UTRA physical layer aspects, TR 36.814, Release 9, Mar. 2010.

[43] J.-C. Lin, T.-H. Lee, and Y.-T. Su, "Power control algorithm for cellular radio systems," *IET Electronics Letters*, vol. 30, no. 3, pp. 195–197, Feb. 1994.

[44] J. Zander, "Distributed cochannel interference control in cellular radio systems," *IEEE Transactions on Vehicular Technology*, vol. 41, no. 3, pp. 305–311, Aug. 1992.

[45] M. Andersin, Z. Rosberg, and Z. Zander, "Gradual removals in cellular pcs with constrained power control and noise," *IEEE International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC)*, Toronto, ON, 1995, vol. 1, pp. 56–60.

[46] D. P. Palomar and M. Chiang, "A tutorial on decomposition methods for network utility maximization," *IEEE Journal on Selected Areas in Communications*, vol. 24, no. 8, pp. 1439–1451, Aug. 2006.

[47] D. Tse and P. Viswanath, *Fundamentals of wireless communication*, Cambridge, United Kingdom: Cambridge University Press, 2005.

[48] R. D. Yates, “A framework for uplink power control in cellular radio systems”, *IEEE Journal on Selected Areas in Communications*, vol. 13, no. 7, pp. 1341–1347, Sep. 1995.

[49] Matlab, MathWorks. Inc., Natick, MA, http://www.mathworks.com.

[50] CVX: Matlab software for disciplined convex programming, version 2.0, CVX Research, Inc., http://cvxr.com/cvx.

[51] Y. Lin, L. Shao, Z. Zhu, Q. Wang, and R. K. Sabhikhi, “Wireless network cloud: architecture and system Requirements,” *IBM Journal of Research and Development*, vol. 54, no. 1, pp. 4:1–4:12, Jan./Feb. 2010.

[52] C-RAN: the road towards green RAN, White Paper, Version 2.5, China Mobile Research Institute, Beijing, China, Oct. 2011.

[53] J. Wu, Z. Zhang, Y. Hong, and Y. Wen, “Cloud radio Access network (C-RAN): a primer,” *IEEE Network*, vol. 29, no. 1, pp. 35–41, Jan. 2015.

[54] M. Peng, Y. Li, J. Jiang, J. Li, and C. Wang, “Heterogeneous cloud radio access networks: a new perspective for enhancing spectral and energy Efficiencies,” *IEEE Wireless Communications*, vol. 21, no. 6, Dec. 2014.

[55] A. D. Domenico, E. C. Strinati, and A. Capone, “Enabling green cellular networks: a survey and outlook,” *Elsevier Computer Communications*, vol. 37, pp. 5–24, Jan. 2014.

[56] C.-C. Hsu, J. M. Chang, and Y.-W. Chen, “Joint optimization for cell configuration and offloading in heterogeneous networks,” in *IEEE International Conference on Computer Communications (INFOCOM)*, San Francisco, CA, 2016, pp. 1–9.

[57] X. Kang, Y.-K. Chia, S. Sun, and H. F. Chong, "Mobile data offloading through a third-party Wifi Access Point: An Operator's Perspective," *IEEE Transactions on Wireless Communications*, vol. 13, no. 10, pp. 5340–5351, Oct. 2014.

[58] S. Li, J. Huang, and S.-Y. R. Li, "Revenue maximization for communication networks with Usage-Based Pricing," in *IEEE Global Communications Conference (GLOBECOM)*, Honolulu, HI, 2009, pp. 1–6.

[59] X. Wang, K. Wang, S. Wu, S. Di, K. Yang, and H. Jin, "Dynamic resource scheduling in could radio access network with mobile cloud computing," in *IEEE/ACM International Symposium on Quality of Service (IWQoS)*, Beijing, China, 2016, pp. 1–6.

[60] S. Luo, R. Zhang, and T. J. Lim, "Downlink and uplink energy minimization through user association and beamforming in C-RAN," *IEEE Transactions on Wireless Communications*, vol. 14, no. 1, pp. 494–508, Jan. 2015.

[61] D. Pompili, A. Hajisami, and H. Viswanathan, "Dynamic provisioning and allocation in cloud radio access networks (C-RANs)," *Elsevier Ad Hoc Networks*, vol. 30, pp. 128–143, Jul. 2015.

[62] C. W. Patterson, A. B. MacKenzie, and S. Glisic, "An economic model of subscriber offloading between mobile network operators and WLAN operators," in *International Symposium on Modeling and Optimization in Mobile, Ad Hoc and Wireless Networks (WiOpt)*, Hammamet, 2014, pp. 444–451.

[63] K. Zhu, E. Hossain, and D. Niyato, "Pricing, spectrum sharing, and service selection in two-tier small cell networks: a hierarchical dynamic game approach," *IEEE Transactions on Mobile Computing*, vol. 13, no. 8, pp. 1843–1856, Jul. 2013.

[64] M. Y. Lyazidi, N. Aitsaadi, and R. Langar, "Dynamic resource allocation for cloud-RAN in LTE with real-time BBU/RRH sssignment," in *IEEE International Conference on Communications (ICC)*, Kuala Lumpur, Malaysia, 2016, pp. 1–6.

[65] L. Feng, W. Li, P. Yu, and X. Qiu, "An rnhanced OFDM resource allocation algorithm in C-RAN based 5G public safety network," *Hindawi Mobile Information Systems*, vol. 2016, 2016. doi:10.1155/2016/9586287.

[66] J. Tang, W. P. Tay, and T. Q. S. Quek, "Cross-layer resource allocation with elastic service scaling in cloud radio access network," *IEEE Transactions on Wireless Communications*, vol. 14, no. 9, pp. 5068–5081, Sept. 2015.

[67] S. Gu, Z. Li, C. Wu, H. Zhang, "Virtualized resource sharing in cloud radio access networks through truthful mechanism," *IEEE Transactions on Communications*, vol. 65, no. 3, pp. 1105–1118, Dec. 2016.

[68] W. Zhao and S. Wang, "Traffic density-based RRH selection for power saving in C-RAN," vol. 34, no. 12, pp. 3157–3167, Dec. 2016.

[69] Y. Cai, F. R. Yu, and S. Bu, "Cloud radio access networks (C-RANs) in mobile cloud computing systems," in *IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*, Toronto, ON, 2014, pp. 369–374.

[70] Y. Cai, F. R. Yu, and S. Bu, "Dynamic operations of cloud radio access networks (C-RANs) for mobile cloud computing systems," *IEEE Transactions on Vehicular Technology*, vol. 65, no. 3, pp. 1536–1548, Mar. 2016.

[71] L. Mashayekhy, M. M. Nejad, and D, Grosu, "Physical machine resource management in clouds: a mechanism design approach," *IEEE Transactions on Cloud Computing*, vol. 3, no. 3, pp. 247–260, Jul./Sept. 2015.

[72] S. Chaisiri, B.-S. Lee, and D. Niyato, “Optimization of resource provisioning cost in cloud computing,” *IEEE Transactions on Services Computing*, vol. 5, no. 2, pp. 164–177, Apr.-Jun. 2012.

[73] K. Guo, M. Sheng, J. Tang, T. Q. S. Quek, and Z. Qiu, “Exploting hybrid clustering and computation provisioning for green C-RAN,” *IEEE Journal on Selected Areas in Communications*, vol. 34, no. 12, pp. 4063–4076, Dec. 2016.

[74] J. Li, M. Peng, Y. Yu, and Z. Ding, “Energy-efficient joint congestion control and resource optimization in heterogeneous cloud radio access networks,” *IEEE Transactions on Vehicular Technology*, vol. 65, no. 12, pp. 9873–9887, Dec. 2016.

[75] M. Peng, Y. Yu, H. Xiang, and H. V. Poor, “Energy-efficient resource allocation optimization for multimedia heterogeneous cloud radio access networks,” *IEEE Transactions on Multimedia*, vol. 18, no. 5, pp. 879–892, May 2016.

[76] J. R. Birge and F. Louveaux, *Introduction to stochastic programming*, ser. Operations Research and Financial Engineering, Springer, 2011.

[77] A. J. King and S. W. Wallace, *Modeling with stochastic programming*, ser. Operations Research and Financial Engineering, Springer, 2012.

[78] M. Peng, C. Wang, V. Lau, and H. V. Poor, “Fronthaul-constrained cloud radio access networks: insights and challenges,” *IEEE Wireless Communications*, vol. 22, no. 2, pp. 152–160, Apr. 2015.

[79] G. C. Pflug, “Scenario tree generation for multiperiod financial optimization by optimal discretization,” *Mathematical Programming*, ser. B, vol. 89, no. 2, pp. 251–271, 2001.

[80] Z. Shen, J. Andrews, and B. Evans, “Adaptive resource allocation in multiuser OFDM systems with proportional rate constraints,” *IEEE Transactions on Wireless Communications*, vol. 4, no. 6, pp. 2726–2737, Nov. 2005.

[81] M. Tao, Y.-C. Liang, and F. Zhang, “Resource allocation for delay differentiated traffic in multiuser OFDM systems,” *IEEE Transactions on Wireless Communications*, vol. 7, no. 6, pp. 2190–2201, Jun. 2008.

[82] Amazon, EC2, https://aws.amazon.com/ec2, 2017.

[83] AT&T, https://www.att.com/.

[84] R. Jain, D.-M. Chiu, and W. Hawe, “A Quantitative Measure of Fairness and Discrimination for Resource Allocation in Shared Computer System,” Digital Equipment Corp., Tech. Rep., Hudson, MA, DEC-TR-301, Sept. 1984.

[85] T. Kim and J. M. Chang, “Enhanced Power Saving Mechanism for Large-Scale 802.11ah Wireless Sensor Networks,” *IEEE Transactions on Green Communications and Networking*, vol. 1, no. 4, pp. 516–527, Dec. 2017.

[86] L. D. Xu, W. He, and S. Li, “Internet of Things in Industries: A Survey,” *IEEE Transactions on Industrial Informatics*, vol. 10, no. 4, pp. 2233–2243, Nov. 2014.

[87] A. Gluhak, S. Krco, M. Nati, D. Pfisterer, N. Mitton, and T. Razafindralambo, “A Survey on Facilities for Experimental Internet of Things Research,” *IEEE Communications Magazine*, vol. 49, no. 11, pp. 58–67, Nov. 2011.

[88] C. Perera, C. H. Liu, S. Jayawardena, and M. Chen, “A Survey on Internet of Things From Industrial Market Perspective,” *IEEE Access*, vol. 2, pp. 1660–1679, Jan. 2014.

[89] C. Perera, A. Zaslavsky, P. Christen, and D. Georgakopoulos, “Context Aware Computing for The Internet of Things: A Survey,” *IEEE Communications Surveys & Tutorials*, vol. 16, no. 1, pp. 414–454, May 2014.

[90] A. Dohr, R. Modre-Osprian, M. Drobics, D. Hayn, G. Schreier, “The Internet of Things for Ambient Assisted Living,” *International Conference on Information Technology: New Generations*, Las Vegas, NV, 2010, pp. 804–809.

[91] T. Adame, A. Bel, R. Bellalta, J. Barcelo, and M. Oliver, “IEEE 802.11ah: The WiFi Approach for M2M Communications,” *IEEE Wireless Communications*, vol. 21, no. 6, pp. 144–152, Dec. 2014.

[92] “Potential Compromise for 802.11ah: Use Case Document,” IEEE 802.11-11/0457r0, Mar. 2011.

[93] M. Park, “IEEE 802.11ah: Sub-1-GHz License-Exempt Operation for the Internet of Things,” *IEEE Communications Magazine*, vol. 53, no. 9, pp. 145–151, Sep. 2015.

[94] S. Aust, R. V. Prasad, and I. G. M. M. Niemegeers, “IEEE 802.11ah: Advantages in standards and further challenges for sub 1 GHz Wi-Fi,” *IEEE International Conference on Communications (ICC)*, Ottawa, ON, 2012, pp. 6885–6889.

[95] O. Raeesi, J. Pirskanen, A. Hazmi, T. Levanen, and M. Valkama, “Performance Evaluation of IEEE 802.11ah and its Restricted Access Window Mechanism,” *IEEE International Conference on Communications Workshops (ICC)*, Sydney, NSW, 2014, pp. 460–466.

[96] E. Khorov, A. Krotov, and A. Lyakhov, “Modeling Machine Type Communication in IEEE 802.11ah Networks,” *IEEE International Conference on Communication Workshop (ICCW)*, London, England, 2015, pp. 1149–1154.

[97] L. Tian, J. Famaey, and S. Latre, "Evaluation of the IEEE 802.11ah Restricted Access Window Mechanism for Dense IoT Networks", *IEEE International Symposium on A World of Wireless, Mobile and Multimedia Networks (WoWMoM)*, Coimbra, Portugal, 2016, pp. 1–9.

[98] L. Zheng, J. Pan, and M. Ni, "Performance Analysis of Grouping Strategy for Dense IEEE 802.11 Networks," *IEEE Global Communications Conference (GLOBECOM)*, Atlanta, GA, 2013, pp. 219–224.

[99] T.-C. Chang, C.-H. Lin, K. C.-J. Lin, and W.-T. Chen, "Load-Balanced Sensor Grouping for IEEE 802.11ah Networks," *IEEE Global Communications Conference (GLOBECOM)*, San Diego, CA, 2015, pp. 1–6.

[100] Y. Zhao, O. N. C. Yilmax, and A. Larmo, "Optimizing M2M Energy Efficiency in IEEE 802.11ah," *IEEE Global Communications Conference Workshops (GC Wkshps)*, San Diego, CA, 2015, pp. 1–6.

[101] E. Khorov, A. Lyakhov, A. Krotov, and A. Guschin, "A Survey on IEEE 802.11ah: An Enabling Networking Technology for Smart Cities," *Elsevier Computer Communications*, vol. 58, pp. 53–69, Mar. 2015.

[102] Y. Zhou, H. Wang, S. Zheng, and Z. Z. Lei, "Advances in IEEE 802.11ah Standardization for Machine-Type Communications in Sub-1GHz WLAN," *IEEE International Conference on Communications Workshops (ICC)*, Budapest, Hungary,, 2013, pp. 1269–1273.

[103] VK Jones and H. Sampath, "Emerging Technologies for WLAN," *IEEE Communications Magazine*, vol. 53, no. 3, pp. 141–149, Mar. 2015.

[104] Y. Wang, Y. Li, K. K. Chai, Y. Chen, and J. Schormans, "Energy-Aware Adaptive Restricted Access Window for IEEE 802.11ah Based Networks," *IEEE Annual International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC)*, Hong Kong, China, 2015, pp. 1211–1215.

[105] Y. Wang, Y. Li, K. K. Chai, Y. Chen, and J. Schormans, "Energy-Aware Adaptive Restricted Access Window for IEEE 802.11ah Based Smart Grid Networks," *IEEE International Conference on Smart Grid Communications (SmartGridComm)*, Miami, FL, 2015, pp. 581–586.

[106] C. W. Park, D. Hwang, and T.-J. Lee, "Enhancement of IEEE 802.11ah MAC for M2M Communications," *IEEE Communications Letters*, vol. 18, no. 7, pp. 1151–1154, July 2014.

[107] A. Argyriou, "Power-Efficient Estimation in IEEE 802.11ah Wireless Sensor Networks with a Cooperative Relay," *IEEE International Conference on Communications (ICC)*, London, England, 2015, pp. 6755–6760.

[108] K. Ogawa, Y. Sangenya, M. Morikura, K. Yamamoto, and T. Sugihara, "IEEE 802.11ah Based M2M Networks Employing Virtual Grouping and Power Saving Methods," *IEEE Vehicular Technology Conference (VTC Fall)*, Las Vegas, NV, 2013, pp. 1–5.

[109] R. P. Liu, G. J. Sutton, and I. B. Collings, "WLAN Power Save with Offset Listen Internal for Machine-to-Machine Communications," *IEEE Transactions on Wireless Communications*, vol. 13, no. 5, pp. 2552–2562, May 2014.

[110] B. Ji, S. Chen, K. Song, C. Li, H. Chen, and Z. Li, "Throughput Enhancement Schemes for IEEE 802.11ah based on Multi-layer Cooperation," *International Wireless Communications and Mobile Computing Conference (IWCMC)*, Dubrovnik, Croatia, 2015, pp. 1112–1116.

[111] M. Qutab-ud-din, A. Hazmi, L. F. D. Carpio, A. Goekceoglu, B. Badihi, P. Amin, A. Larmo, and M. Valkama, “Duty Cycle Challenges of IEEE 802.11ah Networks in M2M and IoT Applications”, *European Wireless Conference*, Oulu, Finland, 2016, pp. 1–7.

[112] B. Badihi, L. F. D. Carpio, P. Amin, A. Larmo, M. Lopez, and D. Denteneer, “Performance Evaluation of IEEE 802.11ah Actuators,” *IEEE Vehicular Technology Conference (VTC Spring)*, Sydney, 2016, pp. 1–5.

[113] M. Anderson, Z. Rosberg, and Z. R. Zander, “Gradual removals in cellular pcs with constrained power control and noise,” *IEEE Annual International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC)*, Toronto, ON, 1995, vol. 1, pp. 56–60.

[114] K. Seong, M. Mohseni, and J. M. Cioffi, “Optimal Resource Allocation for OFDMA Downlink Systems,” *IEEE International Symposium on Information Theory (ISIT)*, Seattle, WA, 2006, pp. 1394–1398.

[115] Q. Wang, M. Hempstead, and W. Yang, “A Realistic Power Consumption Model for Wireless Sensor Network Devices,” *IEEE Communications Society on Sensor and Ad Hoc Communications and Networks (SECON)*, Reston, VA, 2006, pp. 286–295.

[116] T. Kim, D. Qiao, and W. Choi, “Energy-Efficient Scheduling of Internet of Things Devices for Environment Monitoring Applications,” *IEEE International Conference on Communications (ICC)*, Kansas City, MO, May 20–24, 2018.

[117] C.-T. Cheng, C. K. Tse, and F. C. M. Lau, “An energy-aware scheduling scheme for wireless sensor networks,” *IEEE Transactions on Vehicular Technology*, vol. 59, no. 7, pp. 3427–3444, Sept. 2010.

[118] Y. Zhao and J. Wu, “On maximizing the lifetime of wireless sensor networks using virtual backbone scheduling,” *IEEE Transactions on Parallel and Distributed Systems*, vol. 23, no. 8, pp. 1528–1535, Aug. 2012.

[119] T. L. Porta, C. Petrioli, and D. Spenza, “Sensor-mission assignment in wireless sensor networks with energy harvarsting,” *IEEE Communications Society Conference on Sensor, Mech and Ad Hoc Communications and Networks (SECON)* Salt Lake City, UT, 2011, pp. 413–421.

[120] M. P. Johnson, H. Rowaihy, D. Pizzocaro, Amotz Bar-Noy, Stuart Chalmers, Thomas La Porta, and Alune Preece, “Sensor-mission assignment in constrained environments,” *IEEE Transactions on Parallel and Distributed Systems*, vol. 21, no. 11, 2010, pp. 1692–1705, Feb. 2010.

[121] H. Rowaihy, M. Johnson, A. Bar-Noy, T. Brown, and T. L. Porta, “Assigning sensors to competing missions,” *IEEE IEEE Global Communications Conference (GLOBECOM)*, New Orleans, LA, 2008, pp. 1–6.

[122] B. Yener, M. Magdon-Ismail, and F. Sivrikaya, “Joint problem of power optimal connectivity and coverage in wireless sensor networks,” *Wireless Networks*, vol. 13, no. 4, pp. 537–550, Aug. 2007.

[123] C. Liu, K. Wu, and J. Pei, “An energy-efficient data collection framework for wireless sensor networks by exploiting spatiotemporal correlation,” *IEEE Transactions on Parallel and Distributed Systems*, vol. 18, no. 7, pp. 1010–1023, Jul. 2007.

[124] H. Jiang, S. Jin, and C. Wang, “Prediction or not? An energy-efficient framework for clustering-based data collection in wireless sensor networks,” *IEEE Transactions on Parallel and Distributed Systems*, vol. 22, no. 6, pp. 1064–1071, Jun. 2011.

[125] Y. Yao and G. B. Ginnakis, "Energy-efficient scheduling protocols for wireless sensor networks," *IEEE International Conference on Communications (ICC)*, Seoul, Korea, 2005, pp. 2759–2763.

[126] A. Adulyasas, Z. Sun, and N. Wang, "Connected coverage optimization for sensor scheduling in wireless sensor networks," *IEEE Sensors J.*, vol. 15, no. 7, pp. 3877–3892, Jul. 2015.

[127] Q. Cui, H. Wang, P. Hu, X. Tao, P. Zhang, J. Hamalainen, and L. Xia, "Evolution of Limited-Feedback CoMP Systems from 4G to 5G: CoMP Features and Limited-Feedback Approaches," *IEEE Vehicular Technology Magazine*, vol. 9, no. 3, pp. 94–103, 2014.

[128] P. Mach and Z. Becvar, "Mobile Edge Computing: A Survey on Architecture and Computation Offloading," vol. 19, no. 3, pp. 1628–1656, Mar. 2017.