A Two-Timescale Approach for Network Slicing in C-RAN

by

He Zhang

B.E., Xi’an Jiaotong University, 2016

A THESIS SUBMITTED IN PARTIAL FULFILMENT OF
THE REQUIREMENTS FOR THE DEGREE OF

MASTER OF APPLIED SCIENCE

in

THE FACULTY OF GRADUATE AND POSTDOCTORAL STUDIES

(Electrical and Computer Engineering)

THE UNIVERSITY OF BRITISH COLUMBIA

(Vancouver)

December 2018

The following individuals certify that they have read, and recommend to the Faculty of Graduate and Postdoctoral Studies for acceptance, the dissertation entitled:

Profit Maximization for Network Slicing in Cloud Radio Access Networks

submitted by He Zhang in partial fulfillment of the requirements for

the degree of Master of Applied Science

in Electrical and Computer Engineering

Examining Committee:

Vincent W. S. Wong, Electrical and Computer Engineering, UBC, Vancouver

Supervisor

Victor C. M. Leung, Electrical and Computer Engineering, UBC, Vancouver

Supervisory Committee Member

Jane Z. Wang, Electrical and Computer Engineering, UBC, Vancouver

Supervisory Committee Member

Abstract

Network slicing is a promising technique for cloud radio access networks (C-RANs). It enables multiple tenants (i.e., service providers) to reserve resources from an infrastructure provider. However, users' mobility and traffic variation result in resource demand uncertainty for resource reservation. Meanwhile, the inaccurate channel state information (CSI) estimation may lead to difficulties in guaranteeing the quality of service (QoS). To this end, we propose a two-timescale resource management scheme for network slicing in C-RAN, aiming at maximizing the profit of a tenant, which is the difference between the revenue from its subscribers and the resource reservation cost. The proposed scheme is under a hierarchical control architecture which includes long timescale resource reservation for a slice and short timescale intra-slice resource allocation. To handle traffic variation, we utilize the statistics of users' traffic. Moreover, to guarantee the QoS under CSI uncertainty, we apply the uncertainty set of CSI for resource allocation among users. We formulate the profit maximization as a two-stage stochastic programming problem. In this problem, long timescale resource reservation for a slice is performed in the first stage with only the statistical knowledge of users' traffic. Given the decision in the first stage, short timescale intra-slice resource allocation is performed in the second stage, which is adaptive to real-time user arrival and departure. To solve the problem, we first transform the stochastic programming problem into a deterministic optimization problem. We then introduce a maximum interference constraint and transform the QoS constraint under CSI uncertainty into linear matrix inequalities. We further apply semidefinite relaxation to transform the problem into a mixed integer nonconvex optimization problem, which can be solved by combining branch-and-bound and primal-relaxed dual techniques. Simulation results show that our proposed scheme can well adapt to traffic variation and CSI uncertainty. It obtains a higher profit when compared with several baseline schemes.

Lay Summary

Network slicing is a promising technique for the fifth generation (5G) wireless systems. It allows multiple service providers to run on top of a shared physical network infrastructure. Meanwhile, cloud radio access network (C-RAN) is a centralized, cloud computing-based architecture for radio access networks to support various types of traffic demand in 5G wireless systems. However, implementing network slicing in C-RAN is faced with critical challenges due to time-varying network conditions and inaccurate knowledge of the conditions. In this thesis, to tackle the aforementioned challenges, we propose a dynamic resource management scheme for network slicing in C-RAN. Simulation results show that our proposed scheme can achieve a better performance when compared with several baseline schemes.

Preface

I hereby declare that I am the author of this thesis. This thesis is an original, unpublished work under the supervision of Professor Vincent W.S. Wong.

Table of Contents

List of Figures

List of Symbols

$\mathbf{X}^{\mathrm{H}}$	Conjugate transpose of matrix $\mathbf{X}$
$\mathrm{Tr}(\mathbf{X})$	Trace of matrix $\mathbf{X}$
$\mathrm{Rank}(\mathbf{X})$	Rank of matrix $\mathbf{X}$
$\mathbb{R}^{m\times n}$	Set of m by n real matrices
$\mathbb{C}^{m\times n}$	Set of m by n complex matrices
$\mathbb{H}^{N}$	Set of N by N Hermitian matrices
$\mathbf{X}\succeq 0$	Matrix $\mathbf{X}$ is positive semidefinite
$\mathbf{I}_n$	An n by n identity matrix
$\mathbf{0}_n$	An n by 1 all zero vector
$\mathbf{O}_n$	An n by n all zero matrix
$\otimes$	Kronecker product
$\mathfrak{R}\{x\}$	Real part of the complex number x

B	Number of RRHs in the coverage area
$\mathcal{B}$	Set of RRHs in the coverage area
b	An index of an RRH in set $\mathcal{B}$
N	Number of sub-channels
W	Bandwidth of each sub-channel
K	Number of long timescale slots
k	An index of a long timescale slot
T	Number of short timescale slots in each long timescale slot
t	An index of a short timescale slot
$\mathcal{T}_k$	Set of indexes of short timescale slots in the long timescale slot k
M	Number of regions in the coverage area
m	An index of a region
$\chi_{k,m}$	Average user arrival rate in region m in long timescale slot k
$\mu_{k,m}$	Average sojourn time that users stay in region m in long timescale slot k
$\mathcal{U}_{t,m}$	Set of users in short timescale slot t in region m

$\mathcal{U}_t$	Set of users in short timescale slot t in the whole coverage area
u	An index of a user
$\boldsymbol{\chi}_k$	Vector of user arrival rate in long timescale slot k
$\boldsymbol{\mu}_k$	Vector of average sojourn time in long timescale slot k for different regions
n_k	Number of reserved sub-channels in long timescale slot k
$p_{k,b}$	Amount of power reserved for RRH $b \in \mathcal{B}$ in long timescale slot k
$\mathbf{p}_k$	Vector of power reserved for all RRHs
$\mathbf{v}_{t,u,b}$	Beamforming vector from RRH b to user u in short timescale slot t
$\mathbf{v}_{t,u}$	Beamforming vector from all RRHs to user u in short timescale slot t
$\mathbf{v}_t$	Beamforming vector from all RRHs to all users in short timescale slot t
$\bar{\mathbf{h}}_{t,u,b}$	Mean channel vector between RRH b and user u in short timescale slot t
$\bar{\mathbf{h}}_{t,u}$	Mean channel vector for user u in short timescale slot t

$\mathbf{h}_{t,u,b}$	Random channel vector between RRH b and user u in short timescale slot t
$\mathbf{h}_{t,u}$	Random channel vector for user u in short timescale slot t
$r_{t,u}$	Data rate of user u in short timescale slot t
σ^2	Noise power
$\mathcal{R}_{t,u}$	CSI uncertainty set for user u in short timescale slot t
$\varepsilon_{t,u}$	Radius of the uncertainty region of the channel vector $\mathbf{h}_{t,u}$
r^{req}	Required data rate
$Y_{t,u}(\mathbf{v}_t)$	Revenue of serving user u in short timescale slot t under the beamforming decision $\mathbf{v}_t$
α	Revenue by offering the service with 1 Mb/s data rate
β	Penalty of failing to serve user u
$C_k(n_k, p_k)$	Reservation cost of a tenant under the resource reservation decision n_k and $\mathbf{p}_k$
c_1	Cost of reserving one sub-channel in a long timescale slot
c_2	Cost of reserving one Walt of power in a long timescale slot
P_b	Maximum power a tenant can reserve for RRH b

$\mathcal{U}_k^{\mathrm{seq}}$	Sequence of user sets over short timescale slots in long timescale slot k
l	An index of a realization of $\mathcal{U}_k^{\mathrm{seq}}$
$\mathcal{L}$	Set of indexes of realizations of $\mathcal{U}_k^{\mathrm{seq}}$
$\mathcal{U}_{k,l}^{\mathrm{seq}}$	l-th realization of $\mathcal{U}_k^{\mathrm{seq}}$
ω_l	Probability of the occurrence of $\mathcal{U}_{k,l}^{\mathrm{seq}}$
$\mathcal{U}_{t,l}$	User set in short timescale slot t in $\mathcal{U}_{k,l}^{\mathrm{seq}}$
$\mathbf{v}_{k,l}^{\mathrm{seq}}$	Beamforming decision for $\mathcal{U}_{k,l}^{\mathrm{seq}}$ over short timescale slots in long timescale slot k
$\mathbf{v}_{t,l}$	Beamforming vector for $\mathcal{U}_{k,l}^{\mathrm{seq}}$ in short timescale slot t
$a_{t,l,u}$	Binary admission control variable, indicating whether the service request of user u in set $\mathcal{U}_{t,l}$ is accepted or not
$\bar{\mathbf{h}}_{t,l,u}$	Mean channel vector for user u in set $\mathcal{U}_{t,l}$
$\mathbf{h}_{t,l,u}$	Random channel vector for user u in set $\mathcal{U}_{t,l}$
$\mathcal{R}_{t,l,u}$	CSI uncertainty set for user u in set $\mathcal{U}_{t,l}$
$\varepsilon_{t,l,u}^2$	Size of the CSI uncertainty set for user u in set $\mathcal{U}_{t,l}$
$\tilde{\varepsilon}_{t,l,u}^2$	Normalized size of the CSI uncertainty set for user u in set $\mathcal{U}_{t,l}$

$\varphi_{t,l,u}$	An auxiliary variable serving as a lower bound of the signal-to-interference-plus noise ratio for user u in set $\mathcal{U}_{t,l}$
I	Maximum interference threshold
$\mathcal{U}_{t,l}^{\text{relax}}$	Set of users whose $a_{t,l,u}$ is relaxed
$d_{t,l,u}$	Value of $a_{t,l,u}$ that has been determined
$\upsilon_{t,l,u}$	An auxiliary variable for the linear matrix inequality constraint for QoS of user u in set $\mathcal{U}_{t,l}$
$\xi_{t,l,u}$	An auxiliary variable for the linear matrix inequality constraint for QoS of user u in set $\mathcal{U}_{t,l}$
$\mathbf{Q}_{t,l,u}$	An auxiliary matrix for the linear matrix inequality constraint for QoS requirement of user u in set $\mathcal{U}_{t,l}$
$\mathbf{V}_{t,l,u}$	Beamforming matrix for user u in set $\mathcal{U}_{t,l}$
$\mathbf{o}_k$	Sequence of optimization variables in long timescale slot k
$\mathbf{B}_b$	An auxiliary matrix for the power constraint at RRH b

List of Acronyms

BBU	Baseband Unit
BS	Base Station
CoMP	Coordinated Multipoint
C-RAN	Cloud Radio Access Network
CSI	Channel State Information
LMI	Linear Matrix Inequality
mMTC	Massive Machine Type Communications
NFV	Network Function Virtualization
QoS	Quality of Service
RAN	Radio Access Network
RRH	Remote Radio Head
SDN	Software-Defined Networking

SDR	Semidefinite Relaxation
SINR	Signal-to-Interference-plus-Noise Ratio
URLLC	Ultra-Reliable Low-Latency Communications
3GPP	Third Generation Partnership Project
5G	Fifth Generation

Acknowledgements

I would like to express my deepest appreciation to my supervisor, Prof. Vincent Wong, for his invaluable guidance that helped me to shape my research, and for his persistent encouragement and patience that made me fulfill my master program.

I want to thank my senior Dr. Yong Zhou and Dr. Hamed Shah-Mansouri, who gave me constructive comments and help to my research work. I would like to thank my friends in the Communications Lab: Hao Ma, Zehua Wang, Jun Zhu, Bojiang Ma, Yanan Sun, Manyou Ma, Xueting Yang, Xuan Luo, and Xiuhua Li who provided me with generous support and great company. Finally, I own my heartfelt gratitude to my beloved parents for their love, support and understanding.

Dedication

This thesis is dedicated to my beloved family.

Chapter 1

Introduction

In this chapter, we first introduce the background of network slicing and cloud radio access networks (C-RANs). We then introduce the resource management for network slicing, followed by a discussion on the motivation and contributions of our work. The structure of the thesis is shown at the end of this chapter.

1.1 Research Background

1.1.1 Overview of Network Slicing

The fifth generation (5G) wireless systems are expected to support diverse types of services and meet the increasing traffic demands from the end users [1,2]. This scenario leads to higher network capital and operating expenditures, as well as higher network resource consumption. To tackle these problems, network slicing is introduced to virtualize the common physical network into several logical end-to-end networks. Each logical end-to-end network is called a *network slice*. As a logical end-to-end network, each slice consists of a part of core network resources, network functions, and radio access network resources. Each slice can be dynamically created, modified, and released by the centralized controller located at the infrastructure provider. The service provider, which is the owner of each network slice, is called a *tenant*. Based on the network slicing paradigm, each tenant, equipped with a local controller, is capable of managing the network slice according to a specific type of service and quality of service (QoS) requirements including data rate, latency, reliability, and security. There are several crucial requirements for network slicing. First, slice orchestration requires a unified and flexible execution environment to run multiple slices. Second, slice isolation requires separation of resources and independent slice management without interference from other slices. Third, optimized topology and resource allocation are needed to achieve service fulfillment assurance.

The key enablers for network slicing include software-defined networking (SDN) and network function virtualization (NFV) [3]. The main idea of SDN is to decouple the forwarding process of data packets in the data plane from the routing process in the control plane, so that network management can be performed by a logical network controller. In this way, flexibility is achieved by allowing simple and efficient network configuration. The OpenFlow [4, 5] standard is one of the first protocols to implement SDN in the core network. For NFV [6], the main idea is to decouple network functions from the physical network equipment and virtualize these network functions into building blocks that may be chained together to create a specific type of communication service.

1.1.2 Overview of Cloud Radio Access Networks (C-RANs)

C-RAN is a novel mobile network architecture for 5G wireless systems [7]. The main idea behind C-RAN is to detach the radio signal transceiver module and baseband signal processing module of conventional base stations (BSs) into two parts. In C-RAN, the baseband signal processing module is moved from BSs to a cloud server, which is referred to as a baseband unit (BBU). Multiple BBUs running on a cloud server can form a BBU pool, offering centralized baseband signal processing with powerful computation capability. Conventional BSs are replaced by light and low-cost remote radio heads (RRHs) with radio signal transmission and reception functions. To enhance the capacity of C-RAN, the coordinated multipoint (CoMP) transmission technique is deployed by which multiple RRHs can coordinate together to serve each user. The group of RRHs serving each user is called an RRH cluster, and the grouping process is called user-centric RRH clustering. Furthermore, by implementing multiple antennas at each RRH, the beamforming technique can be deployed to mitigate interference experienced by each user.

There are two fundamental downlink data transmission strategies for C-RAN, i.e., data-sharing strategy and compression strategy [8]. In the data sharing strategy, the BBU pool sends messages of each user directly to multiple RRHs by fronthaul links. The RRHs locally form the beamforming vector and cooperatively transmit the messages to each user. In the compression strategy, the central processor located at the BBU pool is responsible for user message precoding. Then, a compressed version of the analog beamformed signals is forwarded to RRHs for cooperative transmission.

1.1.3 Resource Management for Network Slicing

According to different types of network resources, network slicing can be categorized into two types: core network slicing that partitions network nodes, links or topologies, and radio access network (RAN) slicing that partitions baseband resources, BSs, radio resources, and transmission power [9]. Each tenant estimates the resource demand from its subscribed users and submits the resource reservation request to the centralized controller. With the received resource reservation requests, the centralized controller performs inter-slice resource virtualization and assigns the physical resources to each slice. Then, the tenant performs intra-slice resource allocation among its subscribed users. From the perspective of a logical centralized controller, inter-slice resource virtualization is performed by the infrastructure provider, who owns the common physical network. The inter-slice resource virtualization, enabled by SDN and NFV, is responsible of assigning common physical network resources to each slice corresponding to the resource reservation request. Meanwhile, since that the resource reservation requests from different tenants may arrive at different time, inter-slice resource virtualization is also responsible of dynamically scheduling the resources to different tenants [10]. From the perspective of each tenant, resource reservation process and intra-slice resource allocation process can be jointly considered. The resource reservation decision made by the tenant should guarantee sufficient resources for intra-slice resource allocation. Meanwhile, intra-slice resource allocation, performed by the local controller at each tenant, should achieve efficient resource utilization, mitigate interference among users, and guarantee QoS of users.

With the development of C-RAN, implementing network slicing in C-RAN has now attracted more attention and is still an open issue. Besides radio resources considered in RAN slicing, resources of RRHs, fronthaul capacity, and BBU pool need to be considered for network slicing in C-RAN. Moreover, user-centric RRH clustering and beamforming can be integrated to enhance network capacity and achieve efficient resource utilization.

1.2 Motivation

Many research works have been conducted on resource management for core network slicing [10–14]. Compared with core network slicing, RAN slicing is faced with new challenges due to time-varying channel conditions, user mobility, and interference.

Conventional approaches mainly consider inter-slice resource virtualization among tenants from the perspective of a centralized controller to achieve fairness among tenants [15–21]. To achieve accurate resource demand estimation and efficient resource utilization, some studies have been conducted on resource reservation for slices and intra-slice resource allocation from the perspective of each tenant [22, 23]. However, these works consider the two processes in a single timescale framework. To achieve real-time adaptation to varying network conditions, the duration of the timescale is designed to be short. In this case, performing resource reservation and intra-slice resource allocation simultaneously may lead to a high computational cost. To tackle this problem, a two-timescale framework can be adopted. In this framework, resource reservation is performed in a long timescale with the estimated resource demand from the slice, and intra-slice resource allocation is performed in a short timescale to achieve adaptation to real-time network conditions. The two-timescale framework is discussed in several works [24, 25]. However, these works neglect the characterization of the profit of each tenant. To achieve profit maximization, each tenant should control the resource reservation cost and increase the revenue obtained from its subscribed users.

In this thesis, we propose a two-timescale resource management scheme for network slicing in C-RAN, aiming at maximizing the profit of the tenant by long timescale resource reservation for the slice and short timescale intra-slice resource allocation among the subscribed users. We consider two major challenges. First, user traffic varies over time, making it difficult to accurately estimate the resource demand for resource reservation. Second, due to fast fading, user mobility, coding error, and delay, the uncertainty of channel state information (CSI) of the subscribed users should be considered during intra-slice resource allocation in order to guarantee the QoS. To tackle these challenges and maximize the profit of the tenant, the interaction between resource reservation and intra-slice resource allocation is considered. The long timescale resource reservation characterizes the statistics of user traffic and ensures that sufficient resources are reserved for intra-slice resource allocation. Meanwhile, the intra-slice resource allocation is adaptive to the arbitrary arrival/departure of users while characterizing the CSI uncertainty to achieve efficient utilization of the reserved resources and guarantee the QoS.

Considering that the profit maximization problem involves user traffic variation as well as the interaction between resource reservation and intra-slice resource allocation, we apply two-stage stochastic programming to formulate our problem. Stochastic programming is a framework for modeling optimization problems that involve random events [26]. In the two-stage stochastic programming problem, the decision in the first stage is made only with the statistical knowledge of the random event. Then, based on the decision in the first stage and a realization of the random event, the decision in the second stage is made. The objective of the two-stage stochastic programming problem is to maximize the expectation of a certain objective function over the random event. Therefore, by modeling long timescale resource reservation as the decision in the first stage and short timescale intra-slice resource allocation as the decision in the second stage, and modeling the user traffic as the random event, we formulate the profit maximization as a two-stage stochastic programming problem. Moreover, since the CSI uncertainty is difficult to be modeled in a probabilistic manner as many factors (e.g., user mobility, coding error, and delay) lead to the uncertainty, for intra-slice resource allocation in the second stage, we apply the uncertainty set to restrict the realizations of CSI. In this way, the resource allocation decision can be made to guarantee the QoS under the CSI uncertainty.

1.3 Contributions

The main contributions of this thesis are summarized as follows:

- We propose a two-timescale resource management scheme to achieve profit maximization for network slicing in C-RAN. By modeling the problem as a two-stage stochastic programming problem, the interaction between resource reservation and intra-slice resource allocation is achieved, and the user traffic variation is characterized.

- We design a profit model for the tenant, which captures the revenue obtained from its subscribed users and the cost of resource reservation. The revenue is modeled as a piecewise function consisting of a reward obtained by guaranteeing the QoS of users and a penalty due to QoS violation. The cost is modeled as a linear function consisting of the sub-channel and power reservation cost. We characterize the QoS under CSI uncertainty by applying the CSI uncertainty set.

- We transform the stochastic programming problem into a deterministic mixed-integer optimization problem by introducing a maximum interference threshold and applying semidefinite relaxation. We combine branch-and-bound and primal-relaxed dual techniques to obtain the suboptimal solution.

- We conduct extensive simulations to evaluate the properties and performance of the proposed scheme. Results show that the proposed scheme can achieve a higher profit when compared with four other baseline schemes.

1.4 Outline of the Thesis

This thesis is organized as follows. In Chapter 2, we introduce the related work. In Chapter 3, we present the two-timescale resource management scheme to achieve profit maximization for network slicing in C-RAN, analyze its properties, and evaluate its performance. Conclusion and future work are given in Chapter 4.

Chapter 2

Related Work

2.1 Network Slicing

2.1.1 Core Network Slicing

Many studies have been conducted on core network slicing based on SDN and NFV. The framework of FlowVisor is introduced in [11], which is implemented as an OpenFlow proxy that intercepts messages between OpenFlow-enabled switches and OpenFlow controllers. FlowVisor is capable of partitioning link bandwidth and flow table in each switch. Many works have been conducted on the standardization of core network slicing [27,28]. The latest version of network slicing standard is Release 15 of the 3rd Generation Partnership Project (3GPP) completed in June 2018. Several key concepts, such as network slice, network slice instance, and lifecycle management of network slice instance, are specified.

Recently, various dynamic and flexible resource management schemes have been proposed for core network slicing [10, 12]. Baumgartner *et al.* in [12] proposed a robust inter-slice resource allocation scheme with the consideration of survivability to protect network slice against network element (nodes/links) failure. In [10], Sciancalepore *et al.* applied the Holt-

Winters forecasting method to predict the future traffic of each slice based on historical records. With the predicted traffic information, a slice selection and scheduling scheme was proposed, aiming at improving network utilization. Besides resource management schemes for core network slicing to achieve the resource efficiency and utility maximization, there are also other works discussing the reconfiguration of network slicing [13, 14]. Paris *et al.* in [13] addressed the problem that frequent flow reconfigurations for network slicing may cause QoS violation. To tackle this problem, they proposed a control policy to minimize the flow allocation cost while achieving the adaptation to varying network conditions. Pellegrini *et al.* in [14] proposed a learning algorithm to perform optimal online flow segmentation, which can achieve minimum signaling over the control channel and can track traffic variations over time.

2.1.2 Radio Access Network Slicing

Many studies have been conducted on RAN slicing [29]. Early studies on RAN slicing mainly focus on the guarantee of slice isolation. Ravi *et al.* in [15] proposed a time domain resource partitioning scheme to assign different slices into different time slots. However, this scheme does not consider a multi-cell scenario, in which multi-cell interference is a critical problem as slices may share the same spectrum in different cells. To tackle this problem, Gudipati *et al.* in [16] proposed the concept of SoftRAN, which defines a *virtual big-base station* that is comprised of a central controller and a group of geographically close BSs. Besides the resource allocation of time-frequency resource blocks, the authors further proposed a power allocation scheme, which is performed by a logical centralized controller to mitigate inter-cell interference among users in different slices and guarantee slice isolation. Besides the challenge of slice isolation, partitioning RAN resources into different slices corresponding to different use cases and QoS requirements is also an important problem. Caballero *et al.* in [17] focused on achieving desirable fairness across network slices and users. They formulated an optimization problem for dynamic resource allocation with a weighted proportionally fair objective function. Zhang *et al.* in [18] proposed a mobility management scheme and a joint power and subchannel allocation scheme for RAN slicing to enhance resource efficiency. In [19], Xiang *et al.* designed a hierarchical network slicing architecture, consisting of a centralized orchestration layer and a slice instance layer for resource allocation in fog RAN. In [20], Chen *et al.* proposed a resource pre-allocation scheme for each slice and an intra-slice resource scheduling scheme for users with different priorities to achieve resource efficiency. The aforementioned works on RAN slicing assume that the centralized controller can obtain the perfect knowledge of network conditions. However, the uncertainty of network conditions may exist. Zheng *et al.* in [21] proposed a delay-optimal radio resource scheduling scheme with stochastic learning. They applied partially observed Markov decision process to characterize the uncertainty of channel conditions and user traffic for the resource scheduling scheme design.

Instead of considering resource management from the perspective of a centralized controller, some works have been conducted to design the resource reservation and intra-slice resource allocation from the perspective of each tenant. For example, in [23], Zhu *et al.* proposed a hierarchical combinatorial auction mechanism for resource management, in which each tenant submits its bid to the centralized controller for a certain amount of resources, and executes an auction to allocate the reserved resources to its subscribed users. Caballero *et al.* in [22] formulated a network slicing game in which each tenant takes into account the resource demand estimation of other tenants to make a resource reservation decision so as to maximize its user utility. Besides performing resource reservation and intra-slice resource allocation in a single timescale, a two-timescale framework is also considered. In this framework, resource reservation is performed in a long timescale with the predicted resource demands from the slice, and intra-slice resource allocation is performed in a short timescale to achieve the adaptation to real-time network condition variations. Zhang *et al.* in [24] proposed a static spectrum reservation and dynamic resource requesting scheme for each tenant to maximize the aggregate utility of users. In [25], Chen *et al.* designed a resource pre-allocation over a long timescale and intra-slice resource scheduling over a short timescale for resource efficiency maximization.

2.2 C-RAN

Beamforming and user-centric RRH clustering are two major topics in C-RAN. Shi *et al.* in [30] proposed a multi-stage scheme for network power minimization. They separated group sparse beamforming and RRH clustering into different stages. Liu *et al.* in [31] proposed a two-timescale RRH clustering and beamforming scheme to achieve the trade-off between the average weighted sum rate and implementation cost. User-centric RRH clustering is performed in a long timescale and beamforming is performed in a short timescale. Some other works design joint RRH clustering and beamforming. Dai *et al.* in [32] proposed a scheme to perform joint sparse beamforming and user-centric RRH clustering by formulating a zero-norm problem, aiming at maximizing the network utility. Wang *et al.* in [33] proposed a robust beamforming scheme for C-RAN to maximize the network utility. They characterized the QoS under CSI uncertainty by applying the uncertainty set and S-procedure.

2.3 Network Slicing for C-RAN

Research on network slicing for C-RAN is now attracting more attention. Lee *et al.* in [34] proposed a dynamic end-to-end network slicing scheme for heterogeneous C-RAN, which allocates baseband resources, fronthaul/backhaul capacities, and radio resources to multiple tenants with different priorities, aiming at achieving high network throughput while guaranteeing fairness. Costanzo *et al.* in [35] proposed a prototype for network slicing in C-RAN based on Open Air Interface platform and FlexRAN SDN controller to handle the creation and configuration of network slices. Ezzaouia *et al.* in [36] focused on slicing the BBU pool to establish a logical mapping between the BBU pool and RRHs.

2.4 Two-Timescale Resource Management and Profit Maximization

The two-timescale resource management framework is widely used in 5G wireless systems. Liu *et al.* in [31] proposed a two-timescale resource management framework for C-RAN, in which RRH clustering is performed in a long timescale and beamforming is performed in a short timescale. Niu *et al.* in [37] proposed a dynamic resource sharing mechanism among multiple tenants in C-RAN, in which a global resource allocation process is performed in a long timescale and multiple local resource allocation processes are performed in a short timescale. Gao *et al.* in [38] proposed a two-timescale approach for profit maximization in a cloud transcoding system by performing resource provisioning and task scheduling.

Chapter 3

Two-Timescale Resource Management for Network Slicing in C-RAN

3.1 System Model and Problem Formulation

3.1.1 Architecture of Network Slicing in C-RAN

We consider network slicing in a CoMP based C-RAN system. In this system, multiple baseband signal processing modules are located at a BBU pool. The RRHs are composed of radio signal transceivers and are connected to the BBU pool via optical fibers. Each RRH is equipped with multiple antennas. Each user is equipped with a single antenna. The CoMP framework enables each user to be served by multiple RRHs, which form an RRH cluster. Meanwhile, the beamforming operation is designed for antennas to mitigate interference. We apply the data-sharing strategy for downlink data transmission. We denote the set of RRHs in the coverage area as $\mathcal{B} = \{1, 2, \ldots, B\}$. Each RRH is equipped with A antennas. There are N sub-channels, each with bandwidth W. Network slicing is implemented in this CoMP based C-RAN. Each slice corresponds to a virtual network with partial network resources and network functions. Each slice is owned by a tenant (i.e., service provider) to support a specific type of service. In this thesis, we assume that each tenant owns a single slice. Our proposed framework can be easily extended to the scenario where each tenant owns multiple slices.

We consider resource management for network slicing in C-RAN from the perspective of a single tenant. The tenant performs resource reservation for its slice to request radio resources of sub-channels and power from an infrastructure provider, which is the owner of the physical infrastructure of C-RAN. The tenant then performs intra-slice resource allocation to allocate the reserved resources to its subscribed users according to channel conditions and QoS requirements of users.

3.1.2 Two-Timescale Framework

As shown in Fig. 3.1, we divide 24 hours of a day into K long timescale slots (minutes). Each long timescale slot consists of T short timescale slots with the same duration (seconds). Resource reservation is performed over long timescale with the statistical knowledge of user traffic. Intra-slice resource allocation is performed over short timescale under an arbitrary user arrival/departure process. The choice of the duration of each long timescale slot should guarantee that the statistics of user traffic will not change within the long timescale slot. Meanwhile, since that more frequent submissions of resource reservation requests may lead to higher computation and reconfiguration cost of the network, the duration of each long timescale slot should be chosen to avoid high computation and reconfiguration cost. The choice of the duration of each short timescale slot should guarantee that the real-time user traffic variation

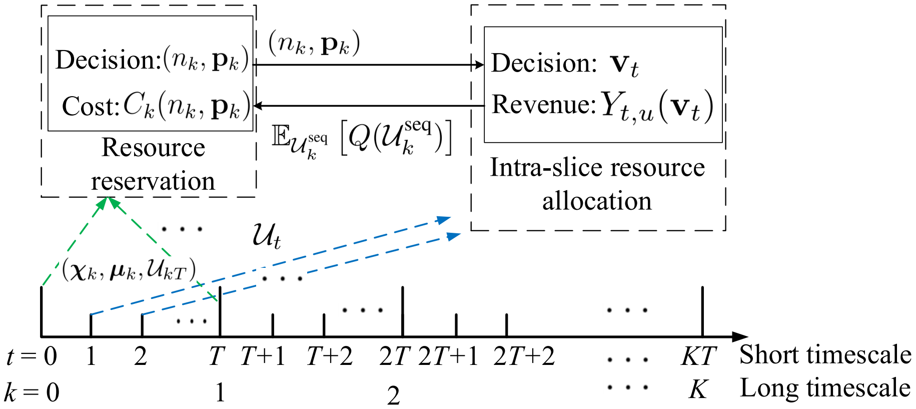
Decision: $(n_k, \mathbf{p}_k)$
Cost: $C_k(n_k, \mathbf{p}_k)$
Resource reservation
$(n_k, \mathbf{p}_k)$
$\mathbb{E}_{\mathcal{U}_k^{\mathrm{seq}}}\left[Q(\mathcal{U}_k^{\mathrm{seq}})\right]$
Decision: $\mathbf{v}_t$
Revenue: $Y_{t,u}(\mathbf{v}_t)$
Intra-slice resource allocation
$\mathcal{U}_t$
$(\boldsymbol{\chi}_k, \boldsymbol{\mu}_k, \mathcal{U}_{kT})$
$t=0$ 1 2 T T+1 T+2 2T 2T+1 2T+2 ... KT Short timescale
$k=0$ 1 2 ... K Long timescale

Figure 3.1: Two-timescale framework with long timescale resource reservation and short timescale intra-slice resource allocation.

can be captured so that the intra-slice resource allocation can be adaptive to arbitrary user arrival and departure. In this thesis, we assume that the durations of each long timescale slot and short timescale slot are predetermined and do not change over time. The explanations of notations in Fig. 3.1 will be given in the following part of this section.

User Traffic Model

In this thesis, we consider the scenario where users arbitrarily arrive and leave the system. In the coverage area of C-RAN, different regions may have different statistics of user traffic. To address this issue, we divide the network coverage area into M disjoint regions, according to the density of user distribution [39]. Within the long timescale slot $k = 0, 1, \ldots, K-1$, in region $m = 1, \ldots, M$, we assume that the arrival of users follows a general distribution with an average user arrival rate of $\chi_{k,m}$ (number of arrived users per short timescale slot). The duration that a user stays in region m, called the *sojourn time*, is a random variable. It follows a general distribution with mean $\mu_{k,m}$, the unit of which is a short timescale slot.

Within a long timescale slot k, we denote the set of users in short timescale slot t as $\mathcal{U}_t = \bigcup_{m=1}^{M} \mathcal{U}_{t,m}$, where $\mathcal{U}_{t,m}$ is the set of users in region m and $t \in \mathcal{T}_k = \{kT, \dots, (k+1)T - 1\}$. Users in set $\mathcal{U}_{t,m}$ are assumed to be uniformly distributed in region m.

Based on the user traffic model, the arrival and departure process of users can be depicted. Within the long timescale slot k, in each short timescale slot t, in each region m, there will be a random number of new user arrivals following the general user arrival distribution with average user arrival rate $\chi_{k,m}$. Each user stays in the region with a random sojourn time. After the sojourn time, the user will leave the system. Since that general distributions for both the arrival of users and the sojourn time are assumed, we can design a resource management scheme that can be applicable for different statistical models of user traffic.

Two-Timescale Resource Management

At the beginning of a long timescale slot k, the tenant obtains the knowledge of user arrival rate vector $\boldsymbol{\chi}_k = (\chi_{k,1}, \dots, \chi_{k,m}, \dots, \chi_{k,M})$, the average sojourn time vector $\boldsymbol{\mu}_k = (\mu_{k,1}, \dots, \mu_{k,m}, \dots, \mu_{k,M})$, and the user set $\mathcal{U}_{kT}$. The tenant then makes the resource reservation decision by choosing n_k, which is the number of reserved sub-channels, and $\mathbf{p}_k = (p_{k,1}, \dots, p_{k,B})$, in which $p_{k,b}$ is the amount of power reserved for RRH $b \in \mathcal{B}$.

At the beginning of a short timescale slot $t \in \mathcal{T}_k$, given the resource reservation decision n_k and $\mathbf{p}_k$, and an observation of user set $\mathcal{U}_t$, we design a beamforming scheme. For a user $u \in \mathcal{U}_t$, the beamforming decision is denoted as $\mathbf{v}_{t,u} = [\mathbf{v}_{t,u,1}^{\mathrm{H}} \cdots \mathbf{v}_{t,u,B}^{\mathrm{H}}]^{\mathrm{H}} \in \mathbb{C}^{AB \times 1}$, where $\mathbf{v}_{t,u,b} \in \mathbb{C}^{A \times 1}$, $b \in \mathcal{B}$, represents the beamforming vector from RRH b to user u for each sub-channel. Furthermore, based on the user location distribution, the mean channel vector of user $u \in \mathcal{U}_t$ can be estimated as $\bar{\mathbf{h}}_{t,u} = [\bar{\mathbf{h}}_{t,u,1}^{\mathrm{H}} \cdots \bar{\mathbf{h}}_{t,u,B}^{\mathrm{H}}]^{\mathrm{H}} \in \mathbb{C}^{AB \times 1}$, where $\bar{\mathbf{h}}_{t,u,b} \in \mathbb{C}^{A \times 1}$, $b \in \mathcal{B}$, is the mean channel vector between RRH b and user u. Due to user mobility and fast channel fading, the instantaneous channel vector, denoted as $\mathbf{h}_{t,u}$, is a random vector, with mean $\bar{\mathbf{h}}_{t,u}$. Given the beamforming decision vector $\mathbf{v}_t = [\mathbf{v}_{t,1}^{\mathrm{H}} \cdots \mathbf{v}_{t,|\mathcal{U}_t|}^{\mathrm{H}}]^{\mathrm{H}}$ and channel vector $\mathbf{h}_{t,u}$, the data rate of user u in short timescale slot t can be obtained as follows [32]:

$$r_{t,u} = n_k W \log \left(1 + \frac{|\mathbf{h}_{t,u}^{\mathrm{H}} \mathbf{v}_{t,u}|^2}{\sum_{u' \in \mathcal{U}_t \setminus \{u\}} |\mathbf{h}_{t,u}^{\mathrm{H}} \mathbf{v}_{t,u'}|^2 + \sigma^2} \right), \tag{3.1}$$

where σ^2 is the noise power. Since the channel vector $\mathbf{h}_{t,u}$ is a random vector, $r_{t,u}$ is also a random variable.

By designing the sparse beamforming vector $\mathbf{v}_{t,u,b}$ for each user $u \in \mathcal{U}_t$ at each RRH $b \in \mathcal{B}$, the tenant can determine the power allocated to user u at RRH b. Meanwhile, the beamforming vector can also indicate the user-centric RRH clustering decision for each user. We note that when $\mathbf{v}_{t,u,b} = \mathbf{0}_{AB}$, user u is not associated with RRH b. When $\mathbf{v}_{t,u,b} \neq \mathbf{0}_{AB}$, user u is served by RRH b.

In this thesis, we assume that resource reservation and intra-slice resource allocation decisions made by the tenant will not be affected by the decisions of other tenants. We also assume that the infrastructure provider can always satisfy the resource reservation requests from the tenant.

QoS Requirement under CSI Uncertainty

In this thesis, the QoS requirement is the required data rate, denoted as r^{req}. Since only the mean channel vector $\bar{\mathbf{h}}_{t,u}$ can be obtained, we adopt the uncertainty set to capture the CSI uncertainty. In short timescale slot $t \in \mathcal{T}_k$, the CSI uncertainty set of user $u \in \mathcal{U}_t$ is defined as

$$\mathcal{R}_{t,u} \triangleq \{\mathbf{h}_{t,u} \mid (\mathbf{h}_{t,u} - \bar{\mathbf{h}}_{t,u})^{\text{H}}(\mathbf{h}_{t,u} - \bar{\mathbf{h}}_{t,u}) \leq \varepsilon_{t,u}^2\}, \tag{3.2}$$

where $\varepsilon_{t,u}$ is the radius of the uncertainty region of the channel vector $\mathbf{h}_{t,u}$. We denote $\varepsilon_{t,u}^2$ as the size of the CSI uncertainty set $\mathcal{R}_{t,u}$. Then, based on (3.1) and (3.2), the QoS requirement under the CSI uncertainty can be modeled as

$$r_{t,u} \geq r^{\text{req}}, \quad \mathbf{h}_{t,u} \in \mathcal{R}_{t,u}. \tag{3.3}$$

Inequality (3.3) indicates that r^{req} should be satisfied for all the realizations of $\mathbf{h}_{t,u}$ in the CSI uncertainty set $\mathcal{R}_{t,u}$. Therefore, by introducing the CSI uncertainty set, the QoS requirement can be depicted as deterministic constraint (3.3) without the necessity of knowing the statistical knowledge of the channel vector.

Revenue and Cost of a Tenant

One key motivation of the tenant to perform resource reservation and intra-slice resource allocation is to enhance the revenue obtained from the subscribed users while controlling the resource reservation cost so as to maximize the profit. In this section, we design a revenue model and a resource reservation cost model for the tenant.

In a short timescale slot $t \in \mathcal{T}_k$, $k = 0, \dots, K-1$, with the knowledge of $\mathcal{U}_t$ and $\mathbf{v}_t$, the revenue of serving user $u \in \mathcal{U}_t$ is given as

$$Y_{t,u}(\mathbf{v}_t) = \begin{cases} p(\tilde{\varepsilon}_{t,u}) r^{\text{req}} \alpha, & r_{t,u} \geq r^{\text{req}},\ \mathbf{h}_{t,u} \in \mathcal{R}_{t,u} \\ -\beta, & \text{otherwise}, \end{cases} \tag{3.4}$$

where $\tilde{\varepsilon}^2_{t,u} = \frac{\varepsilon^2_{t,u}}{\bar{\mathbf{h}}^{\text{H}}_{t,u} \bar{\mathbf{h}}_{t,u}}$ is the normalized size of CSI uncertainty set $\mathcal{R}_{t,u}$, $p(\tilde{\varepsilon}_{t,u})$ is the probability that the true channel vector is within the CSI uncertainty set, $r^{\text{req}}\alpha$ is the revenue of serving user $u \in \mathcal{U}_t$ if perfect CSI information is obtained, in which α is the revenue obtained by offering the service with 1 Mb/s data rate. We also have β as the penalty of failing to serve user u. According to revenue function (3.4), higher required data rate r^{req} results in higher revenue obtained by the tenant, since that users need to pay more for better service. Meanwhile, satisfying QoS constraint (3.3) is not sufficient to guarantee $r_{t,u} \geq r^{\text{req}}$ with 100%, since that the true realization of channel vector $\mathbf{h}_{t,u}$ may be out of the CSI uncertainty set. Therefore, we introduce the probability $p(\tilde{\varepsilon}_{t,u})$, which is determined by $\tilde{\varepsilon}_{t,u}$ of the CSI uncertainty set. Larger $\tilde{\varepsilon}_{t,u}$ may lead to a higher probability that the true realization of channel vector is included in the uncertainty set. Thus, higher probability of QoS guarantee can be achieved for higher revenue. The probability $p(\tilde{\varepsilon}_{t,u})$ of user u can be summarized from historical channel vector records of users located at the same place of user u.

At the beginning of long timescale slot k, given the resource reservation decisions n_k and

$\mathbf{p}_k$, the cost function can be defined as

$$C_k(n_k, \mathbf{p}_k) = c_1 n_k + \sum_{b \in \mathcal{B}} c_2 p_{k,b}, \tag{3.5}$$

where c_1 and c_2 are the costs of reserving one sub-channel and one Walt of power for one long timescale slot, respectively.

3.1.3 Two-Stage Stochastic Programming for Profit Maximization

The objective of long timescale resource reservation and short timescale intra-slice resource allocation is to maximize the expected profit of a tenant in each long timescale slot. For each long timescale slot $k = 0, 1, \ldots, K-1$, we formulate a two-stage stochastic programming problem. The first stage decision, i.e., resource reservation, is made at the beginning of the long timescale slot k, with only the knowledge of the average user arrival rate vector $\boldsymbol{\chi}_k$, average sojourn time vector $\boldsymbol{\mu}_k$, and user set $\mathcal{U}_{kT}$. We denote $\mathcal{U}_k^{\text{seq}} = (\mathcal{U}_{kT}, \ldots, \mathcal{U}_{(k+1)T-1})$. Then, with the first stage decision and a realization of $\mathcal{U}_k^{\text{seq}}$, the second stage decision, i.e., intra-slice resource allocation, is made over short timescale slot $t \in \mathcal{T}_k$. The problem is formulated as follows:

$$\underset{n_k, \mathbf{p}_k}{\text{maximize}} \quad \mathbb{E}_{\mathcal{U}_k^{\text{seq}}}\left[Q(\mathcal{U}_k^{\text{seq}})\right] - C_k(n_k, \mathbf{p}_k) \tag{3.6a}$$

$$\text{subject to} \quad n_k \in \{0, \ldots, N\}, \tag{3.6b}$$

$$0 \le p_{k,b} \le P_b, \quad b \in \mathcal{B}, \tag{3.6c}$$

where P_b is the maximum power a tenant can reserve for RRH $b \in \mathcal{B}$, $Q(\mathcal{U}_k^{\text{seq}})$ is the optimal revenue obtained by the tenant given the knowledge of $\mathcal{U}_k^{\text{seq}}$, $\mathbb{E}_{\mathcal{U}_k^{\text{seq}}}[Q(\mathcal{U}_k^{\text{seq}})]$ is the expectation of $Q(\mathcal{U}_k^{\text{seq}})$ over all the realizations of $\mathcal{U}_k^{\text{seq}}$, $Q(\mathcal{U}_k^{\text{seq}})$ is the optimal value of the following intra-slice resource allocation problem:

$$\underset{\mathbf{v}_t, t\in\mathcal{T}_k}{\text{maximize}} \sum_{t\in\mathcal{T}_k}\sum_{u\in\mathcal{U}_t} Y_{t,u}(\mathbf{v}_t) \tag{3.7a}$$

$$\text{subject to } n_k \sum_{u\in\mathcal{U}_t} \text{Tr}(\mathbf{v}_{t,u,b}\mathbf{v}_{t,u,b}^{\text{H}}) \leq p_{k,b},\ \ b \in \mathcal{B}, t \in \mathcal{T}_k. \tag{3.7b}$$

Constraint (3.7b) represents the power constraint given the decisions of n_k and $\mathbf{p}_k$ made in the first stage.

By solving problem (3.6), the amount of reserved resources and the corresponding cost are determined, based on which second stage problem (3.7) determines the optimal revenue $Q(\mathcal{U}_k^{\text{seq}})$ by making the beamforming decision $\mathbf{v}_t$. Therefore, by solving the two-stage stochastic programming problem, the expected profit can be maximized.

3.2 Solution for the Profit Maximization Problem

3.2.1 Transformation into a Deterministic Problem

The two-stage stochastic programming problem can not be solved directly due to the expectation of $Q(\mathcal{U}_k^{\text{seq}})$ in problem (3.6). Meanwhile, resource reservation in the first stage and intra-slice resource allocation in the second stage build a hierarchical control architecture. Therefore, we first transform the two-stage stochastic programming problem into a deterministic optimization problem [40]. Based on the traffic model in Section 3.1.2, at the beginning of the long timescale slot k, with the knowledge of $(\boldsymbol{\chi}_k, \boldsymbol{\mu}_k, \mathcal{U}_{kT})$, we can obtain the realizations of the user set sequence $\mathcal{U}_k^{\text{seq}}$. The l-th ($l \in \mathcal{L} = \{1, \ldots, L\}$) realization of $\mathcal{U}_k^{\text{seq}}$ is denoted as $\mathcal{U}_{k,l}^{\text{seq}} = (\mathcal{U}_{kT}, \mathcal{U}_{kT+1,l}, \ldots, \mathcal{U}_{(k+1)T-1,l})$. The corresponding probability of the occurrence of realization $\mathcal{U}_{k,l}^{\text{seq}}$ is denoted as ω_l. The corresponding beamforming decision sequence is denoted as $\mathbf{v}_{k,l}^{\text{seq}} = (\mathbf{v}_{kT,l}, \ldots, \mathbf{v}_{(k+1)T-1,l})$, in which $\mathbf{v}_{t,l} = [\mathbf{v}_{t,l,1}^{\text{H}} \cdots \mathbf{v}_{t,l,u}^{\text{H}} \cdots \mathbf{v}_{t,l,|\mathcal{U}_{t,l}|}^{\text{H}}]^{\text{H}}$, and $\mathbf{v}_{t,l,u} = [\mathbf{v}_{t,l,u,1}^{\text{H}} \cdots \mathbf{v}_{t,l,u,B}^{\text{H}}]^{\text{H}}$, $u \in \mathcal{U}_{t,l}$, $t \in \mathcal{T}_k$, $l \in \mathcal{L}$. Then, the two-stage stochastic programming problem can be transformed into the following problem:

$$\underset{n_k, \mathbf{p}_k, \mathbf{v}_k^{\text{seq}}}{\text{maximize}} \sum_{l=1}^{L} \omega_l \sum_{t \in \mathcal{T}_k} \sum_{u \in \mathcal{U}_{t,l}} Y_{t,u}(\mathbf{v}_{t,l}) - C_k(n_k, \mathbf{p}_k) \tag{3.8a}$$

$$\text{subject to} \;\; n_k \sum_{u \in \mathcal{U}_{t,l}} \text{Tr}(\mathbf{v}_{t,l,u,b} \mathbf{v}_{t,l,u,b}^{\text{H}}) \leq p_{k,b}, \quad b \in \mathcal{B}, t \in \mathcal{T}_k, \, l \in \mathcal{L} \tag{3.8b}$$

$$\text{constraints (3.6b) and (3.6c),}$$

where $\mathbf{v}_k^{\text{seq}} = (\mathbf{v}_{k,1}^{\text{seq}}, \ldots, \mathbf{v}_{k,L}^{\text{seq}})$.

Problem (3.8) cannot be solved directly due to the nonconvexity of $Y_{t,u}(\mathbf{v}_{t,l})$. Based on the discussion in Section 3.1.2, the revenue function (3.4) can be equivalently depicted under a user admission control scenario. For user $u \in \mathcal{U}_{t,l}$, the tenant can obtain the revenue $p(\tilde{\varepsilon}_{t,u}) r^{\text{req}} \alpha$ from the user if the QoS requirement constraint (3.3) is satisfied. The tenant will need to pay a penalty of β if constraint (3.3) is not satisfied. To further save the resources, the tenant will then assign no resources to this user, which indicates that the service request of the user is rejected. In this case, we introduce the user admission control variable $a_{t,l,u} \in \{0, 1\}$ to indicate whether the service request of user u is accepted. Then, for the l-th realization of $\mathcal{U}_k^{\text{seq}}$, the revenue function (3.4) is equivalent to

$$Y_{t,l,u}^{\text{new}}(a_{t,l,u}) = a_{t,l,u} p(\tilde{\varepsilon}_{t,l,u}) r^{\text{req}} \alpha - (1 - a_{t,l,u})\beta, \tag{3.9}$$

with QoS constraint

$$\begin{aligned} n_k W \log\left(1 + \frac{|\mathbf{h}_{t,l,u}^{\text{H}} \mathbf{v}_{t,l,u}|^2}{\sum_{u' \in \mathcal{U}_{t,l} \setminus \{u\}} |\mathbf{h}_{t,l,u}^{\text{H}} \mathbf{v}_{t,l,u'}|^2 + \sigma^2}\right) &\geq a_{t,l,u} r^{\text{req}}, \\ \mathbf{h}_{t,l,u} \in \mathcal{R}_{t,l,u}, u \in \mathcal{U}_{t,l}, t \in \mathcal{T}_k, l \in \mathcal{L}&, \end{aligned} \tag{3.10}$$

where $\mathbf{h}_{t,l,u}$, $\mathcal{R}_{t,l,u}$, and $\tilde{\varepsilon}_{t,l,u}^2$ are the channel vector, CSI uncertainty set, and its normalized size for the l-th realization of $\mathcal{U}_k^{\text{seq}}$, respectively. Then, we reformulate problem (3.8) as follows:

$$\underset{\mathbf{a}_k^{\text{seq}}, n_k, \mathbf{p}_k, \mathbf{v}_k^{\text{seq}}}{\text{maximize}} \quad \sum_{l=1}^{L} \omega_l \sum_{t \in \mathcal{T}_k} \sum_{u \in \mathcal{U}_{t,l}} Y_{t,l,u}^{\text{new}}(a_{t,l,u}) - C_k(n_k, \mathbf{p}_k) \tag{3.11a}$$

$$\text{subject to} \quad a_{t,l,u} \in \{0, 1\}, \;\; u \in \mathcal{U}_{t,l}, \; t \in \mathcal{T}_k, \; l \in \mathcal{L} \tag{3.11b}$$

$$\text{constraints (3.6b), (3.6c), (3.8b), (3.10),}$$

where $\mathbf{a}_k^{\text{seq}} = (\mathbf{a}_{k,1}^{\text{seq}}, \ldots, \mathbf{a}_{k,L}^{\text{seq}})$, $\mathbf{a}_{k,l}^{\text{seq}} = (\mathbf{a}_{kT,l}, \ldots, \mathbf{a}_{(k+1)T-1,l})$, and $\mathbf{a}_{t,l} = (a_{t,l,1}, \ldots, a_{t,l,|\mathcal{U}_{t,l}|})$.

Problem (3.11) is a mixed integer optimization problem due to integer variables $\mathbf{a}_k^{\text{seq}}$ and n_k. We use branch-and-bound technique [41] to determine the optimal solution of $\mathbf{a}_k^{\text{seq}}$. We first relax each integer variable $a_{t,l,u} \in \{0,1\}$ to $0 \leq a_{t,l,u} \leq 1$, and solve the relaxed problem to obtain $n_k, \mathbf{p}_k, \mathbf{v}_k^{\text{seq}}$, and relaxed $\mathbf{a}_k^{\text{seq}}$. We randomly choose a variable $a_{t,l,u} \notin \{0,1\}$, the two new constraints developed from this variable are $a_{t,l,u} = 1$ and $a_{t,l,u} = 0$, forming two child nodes of the current node. We then proceed to the node with the greatest optimal value and apply the same procedure. If there is an integer solution of $\mathbf{a}_k^{\text{seq}}$ with the greatest optimal value among other ending nodes, then the process stops. For the integer variable n_k, we relax it to a continuous variable and obtain the relaxed optimal solution of n_k. Then, we simply compare the optimal profits based on the two integer values of n_k that are most close to the relaxed optimal solution of n_k, and pick the optimal integer solution.

3.2.2 QoS Constraint Approximation and Semidefinite Relaxation

Based on the branch-and-bound technique, we focus on solving problem (3.11) with the relaxation of integer variables at each node. The relaxed optimization problem is still difficult to be solved as QoS constraint (3.10) is nonconvex. To tackle this challenge, we introduce a maximum interference threshold to achieve the QoS constraint approximation. The relaxed problem of (3.11) is formulated as follows:

$$\underset{\boldsymbol{\varphi}_k^{\text{seq}}, \mathbf{a}_k^{\text{seq}}, n_k, \mathbf{p}_k, \mathbf{v}_k^{\text{seq}}}{\text{maximize}} \sum_{l=1}^{L} \omega_l \sum_{t \in \mathcal{T}_k} \sum_{u \in \mathcal{U}_{t,l}} Y_{t,l,u}^{\text{new}}(a_{t,l,u}) - C_k(n_k, \mathbf{p}_k) \tag{3.12a}$$

$$\text{subject to} \quad \varphi_{t,l,u} \leq \frac{|\mathbf{h}_{t,l,u}^{\text{H}} \mathbf{v}_{t,l,u}|^2}{I + \sigma^2}, \qquad \mathbf{h}_{t,l,u} \in \mathcal{R}_{t,l,u}, u \in \mathcal{U}_{t,l}, t \in \mathcal{T}_k, l \in \mathcal{L} \tag{3.12b}$$

$$\sum_{u' \in \mathcal{U}_{t,l} \setminus \{u\}} |\mathbf{h}_{t,l,u}^{\text{H}} \mathbf{v}_{t,l,u'}|^2 \leq I, \quad \mathbf{h}_{t,l,u} \in \mathcal{R}_{t,l,u}, u \in \mathcal{U}_{t,l}, t \in \mathcal{T}_k, l \in \mathcal{L} \tag{3.12c}$$

$$n_k W \log\left(1+\varphi_{t,l,u}\right) \geq a_{t,l,u} r^{\text{req}}, \quad u \in \mathcal{U}_{t,l}, t \in \mathcal{T}_k, l \in \mathcal{L} \tag{3.12d}$$

$$0 \leq a_{t,l,u} \leq 1, u \in \mathcal{U}_{t,l}^{\text{relax}},\ t \in \mathcal{T}_k,\ l \in \mathcal{L} \tag{3.12e}$$

$$a_{t,l,u} = d_{t,l,u},\ u \in \mathcal{U}_{t,l} \backslash \mathcal{U}_{t,l}^{\text{relax}}, t \in \mathcal{T}_k, l \in \mathcal{L} \tag{3.12f}$$

$$0 \leq n_k \leq N, \tag{3.12g}$$

constraints (3.6c) and (3.8b),

where $\varphi_{t,l,u}$ is an auxiliary variable serving as a lower bound of the signal-to-interference-plus-noise ratio (SINR), $\boldsymbol{\varphi}_k^{\text{seq}} = (\boldsymbol{\varphi}_{k,1}^{\text{seq}}, \ldots, \boldsymbol{\varphi}_{k,L}^{\text{seq}})$, $\boldsymbol{\varphi}_{k,l}^{\text{seq}} = (\boldsymbol{\varphi}_{kT,l}, \ldots,\ \boldsymbol{\varphi}_{(k+1)T-1,l})$, $\boldsymbol{\varphi}_{t,l} = (\varphi_{t,l,1}, \ldots, \varphi_{t,l,|\mathcal{U}_{t,l}|})$. I is a predefined maximum interference threshold. The optimal solution of problem (3.12) is required to guarantee that the interference experienced by each user is no larger than threshold I. By introducing $\boldsymbol{\varphi}_k^{\text{seq}}$ and I, the QoS constraint (3.10) is relaxed as constraints (3.12b) (3.12c) and (3.12d) [33]. We also have that $\mathcal{U}_{t,l}^{\text{relax}} \in \mathcal{U}_{t,l}$ is the set of users whose $a_{t,l,u}$ is relaxed at the current node. $d_{t,l,u} \in \{0,1\}$ is the value of $a_{t,l,u}$ that has been determined at the current node, in which $u \in \mathcal{U}_{t,l} \backslash \mathcal{U}_{t,l}^{\text{relax}}$, $l \in \mathcal{L}$, $t \in \mathcal{T}_k$.

Due to the CSI uncertainty, constraints (3.12b) and (3.12c) involves infinite number of constraints, making it difficult to directly solve problem (3.12). To tackle this problem, we apply S-procedure [42] to transform constraints (3.12b) and (3.12c) into finite number of linear matrix inequality (LMI) constraints. The S-procedure is introduced in Lemma 3.1:

Lemma 3.1. *(S-Procedure): Let* $\mathbf{A}_1$, $\mathbf{A}_2 \in \mathbb{H}^N$, $\mathbf{d}_1$, $\mathbf{d}_2 \in \mathbb{C}^{N\times 1}$, *and* y_1 $y_2 \in \mathbb{R}$. *Consider the* *following two quadratic functions of vector* $\mathbf{x} \in \mathbb{C}^{N\times 1}$*:*

$$f_1(\mathbf{x}) = \mathbf{x}^H \mathbf{A}_1 \mathbf{x} + 2\Re\{\mathbf{d}_1 \mathbf{x}\} + y_1, \tag{3.13}$$

$$f_2(\mathbf{x}) = \mathbf{x}^H \mathbf{A}_2 \mathbf{x} + 2\Re\{\mathbf{d}_2 \mathbf{x}\} + y_2. \tag{3.14}$$

The implication $f_1(\mathbf{x}) \leq 0 \Rightarrow f_2(\mathbf{x}) \leq 0$ *holds if and only if there exists a* $\theta \geq 0$ *such that*

$$\theta \begin{bmatrix} \mathbf{A}_1 & \mathbf{d}_1 \\ \mathbf{d}_1^H & y_1 \end{bmatrix} - \begin{bmatrix} \mathbf{A}_2 & \mathbf{d}_2 \\ \mathbf{d}_2^H & y_2 \end{bmatrix} \succeq \mathbf{0}. \tag{3.15}$$

We denote that $\Delta\mathbf{h}_{t,l,u} = \mathbf{h}_{t,l,u} - \bar{\mathbf{h}}_{t,l,u}$. Then, by applying Lemma 3.1 to constraint (3.12b), we obtain the following implication:

$$\begin{aligned} &\Delta\mathbf{h}_{t,l,u}^{\mathrm{H}} \mathbf{I}_{AB} \Delta\mathbf{h}_{t,l,u} + 2\Re\{\mathbf{0}^{\mathrm{H}} \triangle \mathbf{h}_u\} - \varepsilon_{t,l,u}^2 \leq 0 \\ &\Rightarrow -\Delta\mathbf{h}_{t,l,u}^{\mathrm{H}} (\mathbf{v}_{t,l,u}\mathbf{v}_{t,l,u}^{\mathrm{H}}) \Delta\mathbf{h}_u - 2\Re\{(\mathbf{v}_{t,l,u}\mathbf{v}_{t,l,u}^{\mathrm{H}} \bar{\mathbf{h}}_{t,l,u})^{\mathrm{H}} \Delta\mathbf{h}_u\} \\ &\qquad\qquad - \bar{\mathbf{h}}_{t,l,u}^{\mathrm{H}} (\mathbf{v}_{t,l,u}\mathbf{v}_{t,l,u}^{\mathrm{H}}) \bar{\mathbf{h}}_{t,l,u} + \varphi_{t,l,u}(I + \sigma^2) \leq 0, \end{aligned} \tag{3.16}$$

if and only if there exists a $\upsilon_{t,l,u} \geq 0$ such that the following LMI holds:

$$\begin{aligned} &\begin{bmatrix} \upsilon_{t,l,u}\mathbf{I}_{AB} & \mathbf{0}_{AB} \\ \mathbf{0}_{AB}^{\mathrm{H}} & -\varphi_{t,l,u}(I+\sigma^2) - \upsilon_{t,l,u}\varepsilon_{t,l,u}^2 \end{bmatrix} + \mathbf{Q}_{t,l,u}^{\mathrm{H}} \mathbf{V}_{t,l,u} \mathbf{Q}_{t,l,u} \succeq \mathbf{0}, \\ &\qquad\qquad u \in \mathcal{U}_{t,l}, t \in \mathcal{T}_k, l \in \mathcal{L}, \end{aligned} \tag{3.17}$$

where $\mathbf{Q}_{t,l,u} = [\mathbf{I}_{AB}\ \bar{\mathbf{h}}_{t,l,u}]$, $\mathbf{V}_{t,l,u} = \mathbf{v}_{t,l,u}\mathbf{v}_{t,l,u}^{\mathrm{H}}$, $\varepsilon_{t,l,u}^2$ is the size of the CSI uncertainty set for the l-th traffic realization.

Similarly, by applying Lemma 3.1 to constraint (3.12c), we obtain the following implication:

$$\begin{aligned}
&\Delta\mathbf{h}_{t,l,u}^{\mathrm{H}}\mathbf{I}_{AB}\Delta\mathbf{h}_{t,l,u} + 2\Re\{\mathbf{0}^{\mathrm{H}}\triangle\mathbf{h}_u\} - \varepsilon_{t,l,u}^2 \leq 0\\
&\Rightarrow \Delta\mathbf{h}_{t,l,u}^{\mathrm{H}}\left(\textstyle\sum_{u'\in\mathcal{U}_{t,l}\setminus\{u\}}\mathbf{V}_{t,l,u'}\right)\Delta\mathbf{h}_{t,l,u} + 2\mathrm{Re}\left\{\left((\textstyle\sum_{u'\in\mathcal{U}_{t,l}\setminus\{u\}}\mathbf{V}_{t,l,u'})\bar{\mathbf{h}}_{t,l,u}\right)^{\mathrm{H}}\Delta\mathbf{h}_{t,l,u}\right\}\\
&\qquad\qquad\qquad + \bar{\mathbf{h}}_{t,l,u}^{\mathrm{H}}\left(\textstyle\sum_{u'\in\mathcal{U}_{t,l}\setminus\{u\}}\mathbf{V}_{t,l,u'}\right)\bar{\mathbf{h}}_{t,l,u} - I \leq 0, \qquad (3.18)
\end{aligned}$$

if and only if there exists a $\xi_{t,l,u} \geq 0$ such that

$$\begin{bmatrix} \xi_{t,l,u}\mathbf{I}_{AB} & \mathbf{0}_{AB} \\ \mathbf{0}_{AB}^{\mathrm{H}} & I - \xi_{t,l,u}\varepsilon_{t,l,u}^2 \end{bmatrix} - \mathbf{Q}_{t,l,u}^{\mathrm{H}}\left(\textstyle\sum_{u'\in\mathcal{U}_{t,l}\setminus\{u\}}\mathbf{V}_{t,l,u'}\right)\mathbf{Q}_{t,l,u} \succeq \mathbf{0},$$

$$u \in \mathcal{U}_{t,l}, t \in \mathcal{T}_k, l \in \mathcal{L}. \qquad (3.19)$$

Then, problem (3.12) is equivalent to

$$\underset{\mathbf{o}_k}{\text{minimize}} \quad C_k(n_k, \mathbf{p}_k) - \sum_{l=1}^{L}\omega_l \sum_{t\in\mathcal{T}_k}\sum_{u\in\mathcal{U}_{t,l}} Y_{t,l,u}^{\text{new}}(a_{t,l,u}) \qquad (3.20a)$$

$$\text{subject to} \quad \text{constraints (3.6c), (3.12d)−(3.12g), (3.17), (3.19),}$$

$$n_k\sum_{u=1}^{|\mathcal{U}_{t,l}|}\mathrm{Tr}(\mathbf{B}_b^{\mathrm{H}}\mathbf{B}_b\mathbf{V}_{t,l,u}) \leq p_{k,b}, \quad b\in\mathcal{B}, t\in\mathcal{T}_k, l\in\mathcal{L} \qquad (3.20b)$$

$$\upsilon_{t,l,u} \geq 0, \qquad u\in\mathcal{U}_{t,l}, t\in\mathcal{T}_k, l\in\mathcal{L} \qquad (3.20c)$$

$$\xi_{t,l,u} \geq 0, \qquad u \in \mathcal{U}_{t,l}, t \in \mathcal{T}_k, l \in \mathcal{L} \tag{3.20d}$$

$$\mathbf{V}_{t,l,u} \succeq \mathbf{0},\ u \in \mathcal{U}_{t,l}, t \in \mathcal{T}_k, l \in \mathcal{L}, \tag{3.20e}$$

$$\text{Rank}(\mathbf{V}_{t,l,u}) \leq 1,\ u \in \mathcal{U}_{t,l}, t \in \mathcal{T}_k, l \in \mathcal{L}, \tag{3.20f}$$

where $\mathbf{o}_k = (\boldsymbol{\varphi}_k^{\text{seq}}, \boldsymbol{\upsilon}_k^{\text{seq}}, \boldsymbol{\xi}_k^{\text{seq}}, \mathbf{a}_k^{\text{seq}}, n_k, \mathbf{p}_k, \mathbf{V}_k^{\text{seq}})$, $\mathbf{B}_b \triangleq (\mathbf{0}_{b-1}^T, 1, \mathbf{0}_{B-b}^T) \otimes \mathbf{I}_A$, so that $\mathbf{v}_{t,l,u,b} = \mathbf{B}_b \mathbf{v}_{t,l,u}$ and $\text{Tr}(\mathbf{v}_{t,l,u,b}\mathbf{v}_{t,l,u,b}^{\text{H}}) = \text{Tr}(\mathbf{B}_b^{\text{H}}\mathbf{B}_b\mathbf{V}_{t,l,u})$. For constraint (3.20f), $\text{Rank}(\mathbf{V}_{t,l,u}) = 0$ happens when $a_{t,l,u} = 0$, meaning that the service request of user u is rejected and there is no resource assigned to that user.

Problem (3.20) is still nonconvex due to constraint (3.20f). We adopt semidefinite relaxation (SDR) [43] by removing constraint (3.20f) to arrive at a tractable problem, given as:

$$\begin{aligned} \underset{\mathbf{o}_k}{\text{minimize}} \quad & C_k(n_k, \mathbf{p}_k) - \sum_{l=1}^{L} \omega_l \sum_{t \in \mathcal{T}_k} \sum_{u \in \mathcal{U}_{t,l}} Y_{t,l,u}^{\text{new}}(a_{t,l,u}) \\ \text{subject to} \quad & \text{constraints (3.6c), (3.12d)}-\text{(3.12g), (3.17), (3.19),} \\ & \text{(3.20b)}-\text{(3.20e).} \end{aligned} \tag{3.21}$$

For the optimal solution of problem (3.21), if the rank of Hermitian matrix $\mathbf{V}_{t,l,u}$ is no larger than one for all $u \in \mathcal{U}_{t,l}$, $l \in \mathcal{L}$ and $t \in \mathcal{T}_k$, then problems (3.20) and (3.21) have the same optimal solution and the same optimal objective value. Otherwise, the optimal objective value of problem (3.20) serves as the lower bound of the optimal objective value of problem (3.21). The tightness of the SDR in problem (3.21) is revealed in the following theorem:

Theorem 3.1. *Assuming problem (3.21) is feasible, a solution for* $\mathbf{V}_{t,l,u}$*,* $\forall\ u \in \mathcal{U}_{t,l}$*,* $l \in \mathcal{L}$*,*

$t \in \mathcal{T}_k$, can always be constructed to guarantee that the rank of the beamforming matrix is less than or equal to one.

Proof. Please refer to Appendix A for the proof of Theorem 3.1. ■

Theorem 3.1 illustrates that if the optimal beamforming solution of problem (3.21) does not meet constraint (3.20f), we can solve problem (A.1) in Appendix A to obtain the optimal solution for beamforming matrix, denoted as $\bar{\mathbf{V}}_{t,l,u}$, $u \in \mathcal{U}_{t,l}$, $t \in \mathcal{T}_k$, $l \in \mathcal{L}$, which satisfies constraint (3.20f). The rank zero solution of $\bar{\mathbf{V}}_{t,l,u}$ indicates that the service request of user u is rejected, i.e., $a_{t,l,u} = 0$. The rank one solution of $\bar{\mathbf{V}}_{t,l,u}$ indicates that the service request of user u is accepted, i.e., $a_{t,l,u} = 1$. Then, the optimal beamforming vector, denoted as $\bar{\mathbf{v}}_{t,l,u}$, is the principle eigenvector of matrix $\bar{\mathbf{V}}_{t,l,u}$.

3.2.3 Primal-Relaxed Dual Technique

Problem (3.21) is still difficult to be solved directly due to the nonconvexity of constraints (3.12d) and (3.20b). One observation is that by fixing variables n_k and $\mathbf{p}_k$, problem (3.21) is convex with respect to variables $\boldsymbol{\varphi}_k^{\text{seq}}$, $\boldsymbol{\upsilon}_k^{\text{seq}}$, $\boldsymbol{\xi}_k^{\text{seq}}$, $\mathbf{a}_k^{\text{seq}}$, $\mathbf{V}_k^{\text{seq}}$. By fixing variables $\boldsymbol{\varphi}_k^{\text{seq}}$, $\boldsymbol{\upsilon}_k^{\text{seq}}$, $\boldsymbol{\xi}_k^{\text{seq}}$, $\mathbf{a}_k^{\text{seq}}$, $\mathbf{V}_k^{\text{seq}}$, problem (3.21) is linear with respect to n_k and $\mathbf{p}_k$. One classical technique to solve this type of optimization problem is the primal-relaxed dual technique [44]. The key idea of primal-relaxed dual technique is to convert the original problem into primal and relaxed dual subproblems that provide valid upper and lower bounds respectively on the global optimal objective value.

We fix variables $\boldsymbol{\varphi}_k^{\text{seq}}, \boldsymbol{\upsilon}_k^{\text{seq}}, \boldsymbol{\xi}_k^{\text{seq}}, \mathbf{a}_k^{\text{seq}}, \mathbf{V}_k^{\text{seq}}$ and solve the primal problem of (3.21) with respect to n_k and $\mathbf{p}_k$, which is formulated as follows:

$$\begin{aligned} \underset{n_k, \mathbf{p}_k}{\text{minimize}} \quad & C_k(n_k, \mathbf{p}_k) - \sum_{l=1}^{L} \omega_l \sum_{t \in \mathcal{T}_k} \sum_{u \in \mathcal{U}_{t,l}} Y_{t,l,u}^{\text{new}}(a_{t,l,u}) \\ \text{subject to} \quad & \text{constraints (3.6c), (3.12d), (3.12g), (3.20b).} \end{aligned} \tag{3.22}$$

The obtained optimal value is denoted as f^{upper}, serving as an upper bound of problem (3.21). The corresponding solutions of Lagrange multipliers for constraints (3.12d), (3.20b), (3.12g), and (3.6c) are denoted as $\lambda_{1,t,l,u}$, $\lambda_{2,t,l,b}$, (for all $u \in \mathcal{U}_{t,l}$, $b \in \mathcal{B}$, $t \in \mathcal{T}_k$, $l \in \mathcal{L}$), λ_3, λ_4, $\lambda_{5,b}$, $\lambda_{6,b}$ (for all $b \in \mathcal{B}$). We use $\boldsymbol{\lambda}$ as the vector of all Lagrange multipliers.

In order to obtain the relaxed dual problem of problem (3.21), we derive the Lagrangian of problem (3.21) with constraints (3.12d), (3.20b), (3.12g), and (3.6c), given as

$$\begin{aligned} & L(\boldsymbol{\varphi}_k^{\text{seq}}, \mathbf{a}_k^{\text{seq}}, \mathbf{V}_k^{\text{seq}}, n_k, \mathbf{p}_k, \boldsymbol{\lambda}) \\ & = c_1 n_k + \sum_{b \in \mathcal{B}} c_2 p_{k,b} - \sum_{l=1}^{L} \omega_l \left(\sum_{t \in \mathcal{T}_k} \sum_{u \in \mathcal{U}_{t,l}} (a_{t,l,u} p(\tilde{\varepsilon}_{t,l,u}) r^{\text{req}} \alpha - (1 - a_{t,l,u}) \beta) \right) \\ & \quad - n_k \sum_{l=1}^{L} \sum_{t \in \mathcal{T}_k} \sum_{u \in \mathcal{U}_{t,l}} \lambda_{1,t,l,u} W \log(1 + \varphi_{t,l,u}) + \sum_{l=1}^{L} \sum_{t \in \mathcal{T}_k} \sum_{u \in \mathcal{U}_{t,l}} \lambda_{1,t,l,u} a_{t,l,u} r^{\text{req}} \\ & \quad - \sum_{b \in \mathcal{B}} p_{k,b} \sum_{l=1}^{L} \sum_{t \in \mathcal{T}_k} \lambda_{2,t,l,b} + n_k \sum_{l=1}^{L} \sum_{t \in \mathcal{T}_k} \sum_{b \in \mathcal{B}} \lambda_{2,t,l,b} \sum_{u \in \mathcal{U}_{t,l}} \text{Tr}(B_b^{\text{H}} B_b V_{t,l,u}) \\ & \quad - \lambda_3 n_k - \lambda_4 N + \lambda^4 n_k - \sum_{b \in \mathcal{B}} \lambda_{5,b} p_{k,b} - \sum_{b \in \mathcal{B}} \lambda_{6,b} P_b + \sum_{b \in \mathcal{B}} \lambda_{6,b} p_{k,b} \\ & = n_k G_1(\boldsymbol{\varphi}_k^{\text{seq}}, \mathbf{V}_k^{\text{seq}}, \boldsymbol{\lambda}) + G_2(\mathbf{a}_k^{\text{seq}}, \mathbf{p}_k, \boldsymbol{\lambda}), \end{aligned} \tag{3.23}$$

where

$$\begin{aligned} G_1(\boldsymbol{\varphi}_k^{\text{seq}}, \mathbf{V}_k^{\text{seq}}, \boldsymbol{\lambda}) = \quad & c_1 - \sum_{l=1}^{L} \sum_{t \in \mathcal{T}_k} \sum_{u \in \mathcal{U}_{t,l}} \lambda_{1,t,l,u} W \log(1 + \varphi_{t,l,u}) \\ & + \sum_{l=1}^{L} \sum_{t \in \mathcal{T}_k} \sum_{b \in \mathcal{B}} \lambda_{2,t,l,b} \sum_{u \in \mathcal{U}_{t,l}} \text{Tr}(\mathbf{B}_b^{\text{H}} \mathbf{B}_b \mathbf{V}_{t,l,u}) - \lambda_3 + \lambda_4, \end{aligned}$$

and

$$\begin{aligned} G_2(\mathbf{a}_k^{\text{seq}}, \mathbf{p}_k, \boldsymbol{\lambda}) = \quad & \sum_{b \in \mathcal{B}} p_{k,b} \left(c_2 - \sum_{l=1}^{L} \sum_{t \in \mathcal{T}_k} \lambda_{2,t,l,b} - \lambda_{5,b} + \lambda_{6,b} \right) \\ & - \sum_{l=1}^{L} \omega_l \left(\sum_{t \in \mathcal{T}_k} \sum_{u \in \mathcal{U}_{t,l}} \left(a_{t,l,u} p(\tilde{\varepsilon}_{t,l,u}) r^{\text{req}} \alpha - (1 - a_{t,l,u}) \beta \right) \right) \\ & + \sum_{l=1}^{L} \sum_{t \in \mathcal{T}_k} \sum_{u \in \mathcal{U}_{t,l}} \lambda_{1,t,l,u} a_{t,l,u} r^{\text{req}} - \lambda_4 N - \sum_{b \in \mathcal{B}} \lambda_{6,b} P_b. \end{aligned}$$

With the Lagrangian, we further have

$$\begin{aligned} & \inf_{0 \le n_k \le N} L(\boldsymbol{\varphi}_k^{\text{seq}}, \mathbf{a}_k^{\text{seq}}, \mathbf{V}_k^{\text{seq}}, n_k, \mathbf{p}_k, \boldsymbol{\lambda}) = \inf_{0 \le n_k \le N} n_k G_1(\boldsymbol{\varphi}_k^{\text{seq}}, \mathbf{V}_k^{\text{seq}}, \boldsymbol{\lambda}) + G_2(\mathbf{a}_k^{\text{seq}}, \mathbf{p}_k, \boldsymbol{\lambda}) \\ & = \frac{N - \delta N}{2} G_1(\boldsymbol{\varphi}_k^{\text{seq}}, \mathbf{V}_k^{\text{seq}}, \boldsymbol{\lambda}) + G_2(\mathbf{a}_k^{\text{seq}}, \mathbf{p}_k, \boldsymbol{\lambda}), \end{aligned} \tag{3.24}$$

where $\delta \in \{-1, 1\}$ such that $\delta G_1(\boldsymbol{\varphi}_k^{\text{seq}}, \mathbf{V}_k^{\text{seq}}, \boldsymbol{\lambda}) \ge 0$. It indicates that when $G_1(\boldsymbol{\varphi}_k^{\text{seq}}, \mathbf{V}_k^{\text{seq}}, \boldsymbol{\lambda}) \le 0$, δ will be equal to -1, which is equivalent to have $n_k = N$ that achieves the minimization of Lagrangian over n_k. When $G_1(\boldsymbol{\varphi}_k^{\text{seq}}, \mathbf{V}_k^{\text{seq}}, \boldsymbol{\lambda}) \ge 0$, δ will be equal to 1, which is equivalent to have $n_k = 0$ that achieves the minimization of Lagrangian over n_k.

By fixing Lagrange variables $\boldsymbol{\lambda}$ and $\mathbf{p}_k$, based on the analysis in [44], we obtain the relaxed dual problem of problem (3.21) as follows:

$$
\begin{aligned}
\underset{\boldsymbol{\varphi}_k^{\mathrm{seq}},\boldsymbol{\upsilon}_k^{\mathrm{seq}},\boldsymbol{\xi}_k^{\mathrm{seq}},\mathbf{a}_k^{\mathrm{seq}},\mathbf{V}_k^{\mathrm{seq}},\delta}{\text{minimize}} \quad & \frac{N-\delta N}{2} G_1(\boldsymbol{\varphi}_k^{\mathrm{seq}},\mathbf{V}_k^{\mathrm{seq}}) + G_2(\mathbf{a}_k^{\mathrm{seq}}) && \text{(3.25a)} \\
\text{subject to} \quad & \text{constraints (3.12e), (3.12f), (3.17), (3.19), (3.20c)}-\text{(3.20e)} && \\
& \delta G_1(\boldsymbol{\varphi}_k^{\mathrm{seq}},\mathbf{V}_k^{\mathrm{seq}}) \geq 0, && \text{(3.25b)} \\
& \delta \in \{-1, 1\}. && \text{(3.25c)}
\end{aligned}
$$

The optimal value of problem (3.25) is denoted as f^{lower}, serving as a lower bound of problem (3.21). We iteratively solve the primal problem (3.22) and the relaxed dual problem (3.25) until the gap between the upper and lower bounds is below a predetermined threshold.

We present a flow chart to depict the whole process of our problem transformation, as shown in Fig. 3.2. In order to efficiently solve the profit maximization problem, which is originally formulated as a two-stage stochastic programming problem, several transformation and approximation steps should be taken. The two-stage stochastic programming problem consists of problems (3.6) and (3.7). We first transform the problem into deterministic optimization problem (3.8). Due to the nonconvexity of revenue function (3.4), then we transform the revenue function into a linear function with QoS constraint (3.10) by introducing a user admission control variable $a_{t,l,u}$ for each user. Moreover, problem (3.8) can be transformed into problem (3.11). Due to the nonconvexity of QoS constraint (3.10), we introduce a maximum interference threshold I to achieve QoS approximation. We also relax the integer variables to continuous variables and obtain the relaxed optimization problem (3.12).

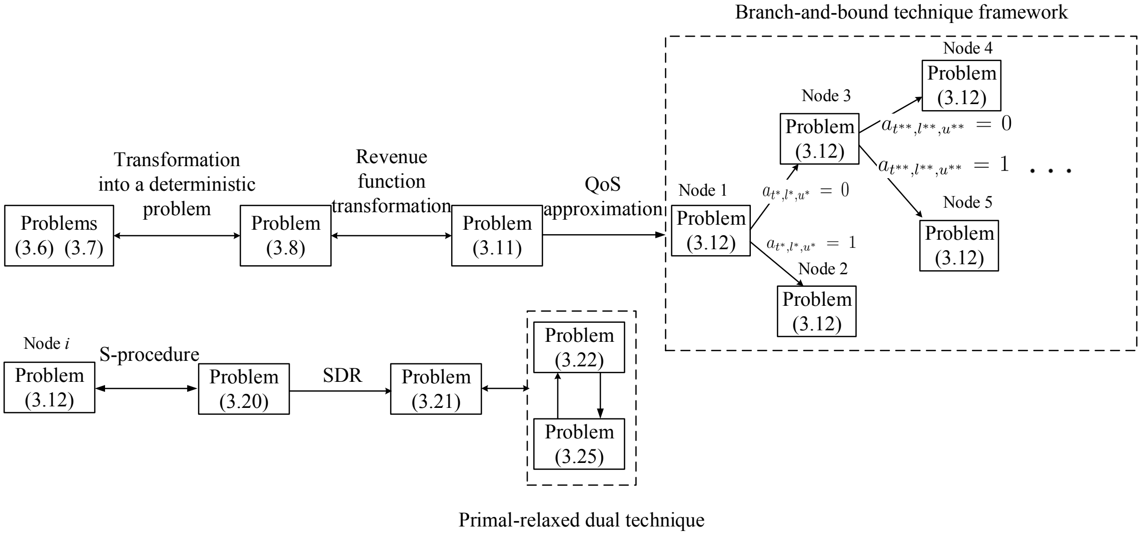

Figure 3.2: The transformation and relaxation steps taken from problems (3.6) and (3.7) to obtain the solutions of profit maximization problem.

Then, we apply the branch-and-bound technique to obtain the optimal integer solution of $a_{t,l,u}$ for each user. In the framework of the branch-and-bound technique, we solve the relaxed optimization problem (3.12) for each node. To solve this problem, we apply S-procedure and SDR to obtain problem (3.21), which can be solved by applying primal-relaxed dual technique. In Fig. 3.2, a bidirectional arrow represents a transformation into a equivalent problem. The unidirectional arrows from problem (3.11) to problem (3.12), and from problem (3.20) to problem (3.21), represent transformations involving approximations.

3.2.4 Joint Resource Reservation and Allocation Algorithm

In this section, we design the algorithm to achieve the two-timescale resource management for network slicing in C-RAN, with the objective of maximizing the profit of the tenant. We first design the algorithm to depict the primal-relaxed dual technique for each node, which is shown in Algorithm 3.1. We then design the global algorithm for resource reservation and

Algorithm 3.1: Primal-Relaxed Dual Technique for Node i

```
1  Input $\mathcal{D}(i)$, $\mathbf{a}_k^{\text{seq}}(i)$, and $\mathcal{U}_{t,l}^{\text{relax}}(i)$, $t \in \mathcal{T}_k$, $l \in \mathcal{L}$.
2  $j := 1$.
3  Initialize $\boldsymbol{\varphi}_k^{\text{seq}}(i,j), \boldsymbol{\upsilon}_k^{\text{seq}}(i,j), \boldsymbol{\xi}_k^{\text{seq}}(i,j), \mathbf{V}_k^{\text{seq}}(i,j)$ subject to constraints (3.17), (3.19),
   (3.20c)$-$(3.20e); Set $\epsilon := 10^{-3}$
4  $\mathbf{a}_k^{\text{seq}}(i,j) := \mathbf{a}_k^{\text{seq}}(i)$.
5  $f^{\text{lower}}(i,j) := -\infty$, $f^{\text{upper}}(i,j) := 0$.
6  while $|f^{\text{upper}}(i,j) - f^{\text{lower}}(i,j)| \geq \epsilon$ do
7      Solve problem (3.22) with fixed $\boldsymbol{\varphi}_k^{\text{seq}}(i,j)$, $\boldsymbol{\upsilon}_k^{\text{seq}}(i,j)$, $\boldsymbol{\xi}_k^{\text{seq}}(i,j)$, $\mathbf{a}_k^{\text{seq}}(i,j)$, $\mathbf{V}_k^{\text{seq}}(i,j)$,
       update $n_k(i,j+1)$, $\mathbf{p}_k(i,j+1)$ and $f^{\text{upper}}(i,j+1)$ with the optimal solutions.
8      Solve relaxed dual problem (3.25) with fixed $\mathbf{p}_k(i,j)$ and dual variables obtained in
       Step 7, with $\mathcal{D}(i)$ and $\mathcal{U}_{t,l}^{\text{relax}}(i)$, $t \in \mathcal{T}_k$, $l \in \mathcal{L}$; update $\boldsymbol{\varphi}_k^{\text{seq}}(i,j+1)$, $\boldsymbol{\upsilon}_k^{\text{seq}}(i,j+1)$,
       $\boldsymbol{\xi}_k^{\text{seq}}(i,j+1)$, $\mathbf{a}_k^{\text{seq}}(i,j+1)$, $\mathbf{V}_k^{\text{seq}}(i,j+1)$, $f^{\text{lower}}(i,j+1)$.
9      $j := j+1$.
10 end
11 Return $f^{\text{upper}}(i,j)$, $f^{\text{lower}}(i,j)$, and optimal solution $\mathbf{o}_k(i,j) := (\boldsymbol{\varphi}_k^{\text{seq}}(i,j)$,$\boldsymbol{\upsilon}_k^{\text{seq}}(i,j)$,
   $\boldsymbol{\xi}_k^{\text{seq}}(i,j)$, $\mathbf{V}_k^{\text{seq}}(i,j)$, $\mathbf{a}_k^{\text{seq}}(i,j)$, $n_k(i,j)$, $\mathbf{p}_k(i,j))$.
```

intra-slice resource allocation in long timescale slot $k = 0, \ldots, K$ based on the branch-and-bound technique and integrate Algorithm 3.1 in the inner iteration. This algorithm is shown in Algorithm 3.2.

In Algorithms 3.1 and 3.2, we introduce set $\mathcal{D}(i)$ at node i to record the determined value $d_{t,l,u}$ and the corresponding index for the user admission control variable $a_{t,l,u}$. In Algorithm 3.2, steps $7-14$ depict the process to calculate the optimal solutions for the two child nodes generated from the last chosen node. Steps $15-18$ depict the process of choosing the ending node with the greatest optimal objective value and initializing the two child nodes. Theoretically, the worst case time complexity of Algorithm 3.2 is dominated by the branch-and-bound technique, and is $\mathcal{O}(2^n)$, where n is the total number of user admission control

Algorithm 3.2: Global Algorithm for Resource Reservation and Intra-Slice Resource Allocation in Long Timescale Slot k

```
Set $i := 1$, in which $i$ represents the index of the node of the branch-and-bound
  technique.
Initialize the admission control decision vector $\mathbf{a}_k^{\text{seq}}(i)$ for the outer iteration by
  randomly assigning a value within $[0,1]$ to each $a_{t,l,u}(i)$, $u \in \mathcal{U}_{t,l}$, $t \in \mathcal{T}_k$, $l \in \mathcal{L}$.
$\mathcal{U}_{t,l}^{\text{relax}}(i) := \mathcal{U}_{t,l}$, $t \in \mathcal{T}_k$, $l \in \mathcal{L}$.
Initialize $\mathcal{D}(i) := \emptyset$ to record the set of $(d_{t,l,u}, t, l, u)$ at the current node.
Initialize $\mathcal{F}_{\text{node}} := \emptyset$ to record the optimal values and solutions of ending nodes and the
  indexes of the nodes.
/* Outer Iteration: Branch-and-Bound technique */
while $\exists\ a_{t,l,u}(i) \notin \{0,1\}$, $\forall u, t, l$, do
    $s := 1$.
    while $s \leq 2$ do
        /* Inner Iteration: Primal-Relaxed Dual Technique */
        Perform Algorithm 3.1, with input $\mathcal{D}(i)$, $\mathbf{a}_k^{\text{seq}}(i)$, and $\mathcal{U}_{t,l}^{\text{relax}}(i)$, $t \in \mathcal{T}_k$, $l \in \mathcal{L}$.
        $f := \frac{f^{\text{upper}(i,j)}+f^{\text{lower}(i,j)}}{2}$.
        $\mathbf{o}_k := \mathbf{o}_k(i,j)$.
        $\mathcal{F}_{\text{node}} := \mathcal{F}_{\text{node}} \bigcup \{(f, \mathbf{o}_k, i)\}$.
        $i := i+1$, $s := s+1$.
    end
    $(f^*, \mathbf{o}^*, i^*) := \arg\min_{f^*}\{\mathcal{F}_{\text{node}}\}$; update $\mathbf{a}_k^{\text{seq}}(i)$ and $\mathbf{a}_k^{\text{seq}}(i+1)$ based on $\mathbf{o}^*$;
      Randomly choose $a_{t^*,l^*,u^*} \notin \{0,1\}$ in $\mathbf{o}^*$.
    $\mathcal{D}(i) := \mathcal{D}(i^*) \bigcup \{(0, t^*, l^*, u^*)\}$, $\mathcal{D}(i+1) := \mathcal{D}(i^*) \bigcup \{(1, t^*, l^*, u^*)\}$.
    $\mathcal{U}_{t^*,l^*}^{\text{relax}}(i) := \mathcal{U}_{t^*,l^*}^{\text{relax}}(i^*) \backslash \{u^*\}$, $\mathcal{U}_{t^*,l^*}^{\text{relax}}(i+1) := \mathcal{U}_{t^*,l^*}^{\text{relax}}(i^*) \backslash \{u^*\}$.
    $\mathcal{F}_{\text{node}} := \mathcal{F}_{\text{node}} \backslash \{(f^*, \mathbf{o}^*, i^*)\}$.
end
Return $-f^*$.
```

variables. However, in practice, the algorithm can run fast as only a small number of nodes are searched before reaching the optimal solutions.

3.3 Performance Evaluation

3.3.1 Simulation Environment and Parameter Setup

The coverage area of the C-RAN network is 300×300 m^2. It is divided into nine regions. Each region is 100×100 m^2 with an RRH at its center. Thus, the number of RRHs is 9. Each RRH is equipped with two antennas. The total bandwidth is 20 MHz, which is divided into 20 sub-channels. The channel model consists of path loss and small scale fading which follows Rayleigh fading. The reference distance for path loss estimation is 2 m. The path loss exponent is 3.6. The mean channel vector $\bar{\mathbf{h}}_{t,u}$ of user $u \in \mathcal{U}_t$ in short timescale slot $t \in \mathcal{T}_k$ is determined by the path loss. The noise power σ^2 is -101 dBm, and the noise of each user follows the zero-mean complex Gaussian distribution with variance σ^2. We set the interference threshold $I = 28\sigma^2$. The duration of each long timescale slot and short timescale slot are 20 minutes and 5 seconds, respectively. The sub-channel reservation cost c_1 is set to be \$$0.05$. The power reservation cost c_2 is set as \$$0.05$. The reward α is \$$0.005$. The penalty β is \$$0.003$. The arrival process of users follows Poisson distribution. The average user arrival rate $\chi_{k,m}$, $m \in \mathcal{M}$ is chosen uniformly within $[\bar{\chi} - \Delta\chi, \bar{\chi} + \Delta\chi]$, $\Delta\chi = 1$ (number of users per short timescale slot). The sojourn time of users follows the uniform distribution within $[2,\ 10]$, the unit of which is a short timescale slot. The normalized size $\tilde{\varepsilon}^2_{t,l,u}$ of CSI uncertainty set is chosen uniformly within $[\bar{\varepsilon}^2 - \Delta\bar{\varepsilon}^2, \bar{\varepsilon}^2 + \Delta\bar{\varepsilon}^2]$, $\Delta\bar{\varepsilon}^2 = \frac{\bar{\varepsilon}^2}{2}$. In our simulation, dividing the coverage area into disjoint regions is only for characterizing different statistics of user traffic in different regions. The RRHs located in different regions are still able to coordinate together to serve each user.

3.3.2 Algorithm Properties

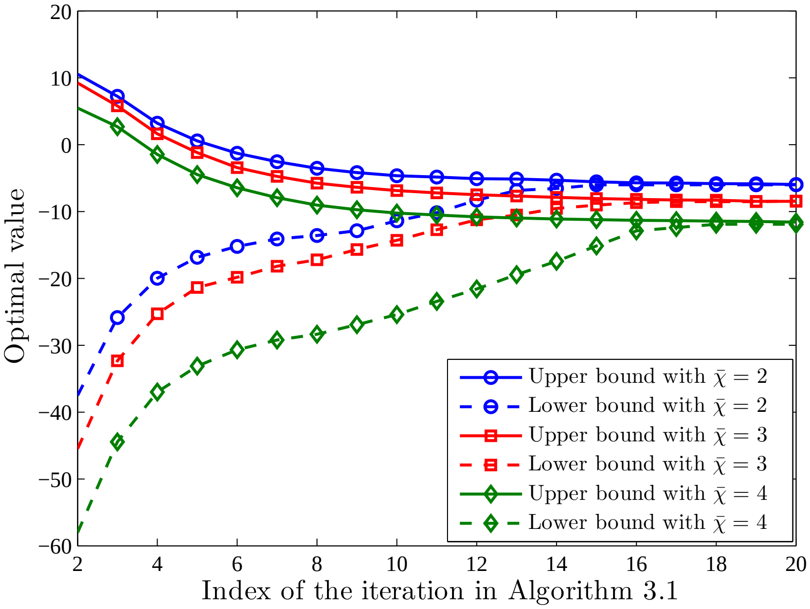

Figure 3.3: Convergence of Algorithm 3.1 with different average user arrival rate $\bar{\chi}$ (number of users per short timescale slot), $r^{\text{req}} = 1.5$ Mb/s, $\bar{\varepsilon}^2 = 0.05$.

In this section, we evaluate the properties of the proposed algorithm. We first conduct simulations to evaluate the impact of user traffic on the convergence of Algorithms 3.1 and 3.2. The simulation results are shown in Figs. 3.3, 3.4 and 3.5. Fig. 3.3 shows the convergence of Algorithm 3.1. Each iteration represents the process of solving problems (3.22) and (3.25) to obtain an upper bound and lower bound. The algorithm converges when the gap between the upper bound and lower bound is smaller than a predetermined threshold. As the average user arrival rate $\bar{\chi}$ (number of arrived users per short timescale slot) increases, the convergence rate becomes slower. This is because that larger $\bar{\chi}$ leads to a larger number of users in the coverage

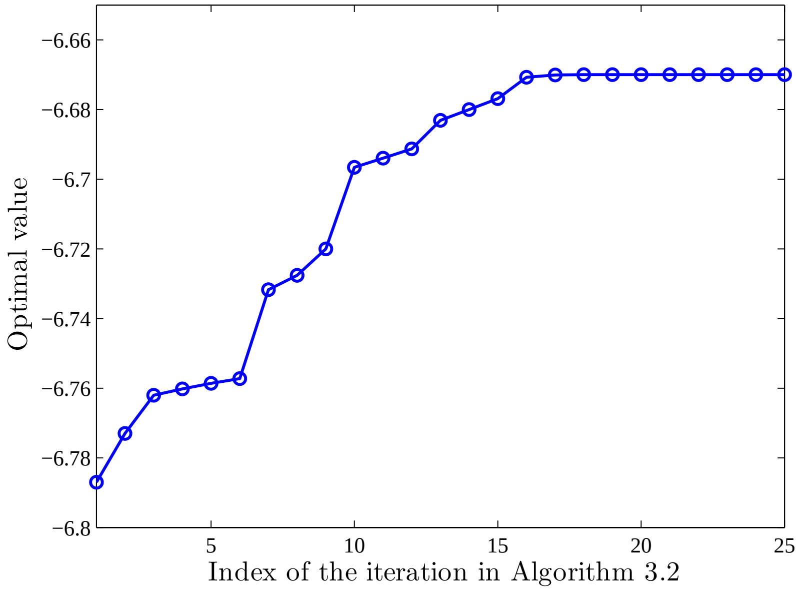

Figure 3.4: Outer iteration convergence of Algorithm 3.2, $r^{\text{req}} = 1.5$ Mb/s, $\bar{\varepsilon}^2 = 0.05$, $\bar{\chi} = 2$ (number of users per short timescale slot).

area, thus a larger number of variables to be solved at each iteration in Algorithm 3.1. However, the difference among convergence rates under different $\bar{\chi}$ is negligible. So we can conclude that the user traffic variation only has a minor impact on the convergence of Algorithm 3.1.

Figs. 3.4 and 3.5 show the outer iteration convergence of Algorithm 3.2 with the average user arrival rate $\bar{\chi} = 2$ (number of arrived user per short timescale slot) and $\bar{\chi} = 4$ (number of arrived user per short timescale slot), respectively. Each iteration consists of obtaining the converged solution of Algorithm 3.1 at the two child nodes generated from the last node, preceding to the node with the greatest optimal objective value, and generating two new child nodes. The algorithm converges when we obtain an integer solution of $\mathbf{a}_k^{\text{seq}}$ with the greatest optimal objective value among all ending nodes. In Fig. 3.5, $\bar{\chi} = 4$, which is larger than $\bar{\chi} = 2$ in Fig. 3.4. So, there is a larger number of users in the system, which leads to a larger number

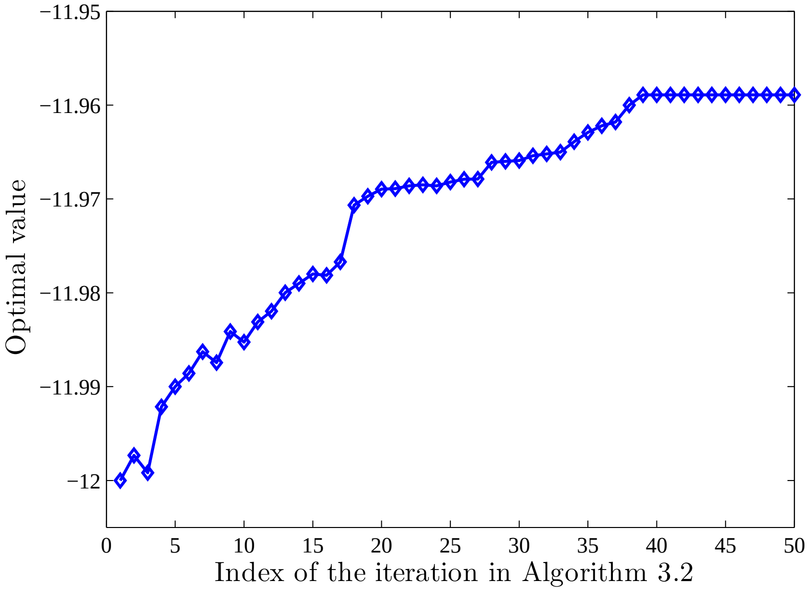

Figure 3.5: Outer iteration convergence of Algorithm 3.2, $r^{\text{req}} = 1.5$ Mb/s, $\bar{\varepsilon}^2 = 0.05$, $\bar{\chi} = 4$ (number of users per short timescale slot).

of user admission control variables of $a_{t,l,u}$. Therefore, for branch-and-bound technique, it takes longer time to find integer solutions for all $a_{t,l,u}$, $u \in \mathcal{U}_{t,l}$, $l \in \mathcal{L}$, $t \in \mathcal{T}_k$. In this case, the convergence rate in Fig. 3.5 is slower than that in Fig. 3.4. However, the convergence is still fast in practice, compared with the theoretical worst case complexity of $\mathcal{O}(2^n)$. This is because that at the first iteration of Algorithm 3.2, $a_{t,l,u}$ for those users with good channel quality are directly assigned to be one. Meanwhile, for those users with really bad channel quality, $a_{t,l,u}$ are directly assigned to be zero. Then, the branch-and-bound technique only needs to justify the optimal integer solutions of $a_{t,l,u}$ for a small number of users.

As discussed in Section 3.2.2, we can not get the exact optimal solution and the optimal profit for two reasons. First, we introduced a maximum interference threshold. Second, we applied semidefinite relaxation. The semidefinite relaxation has been analyzed in Appendix

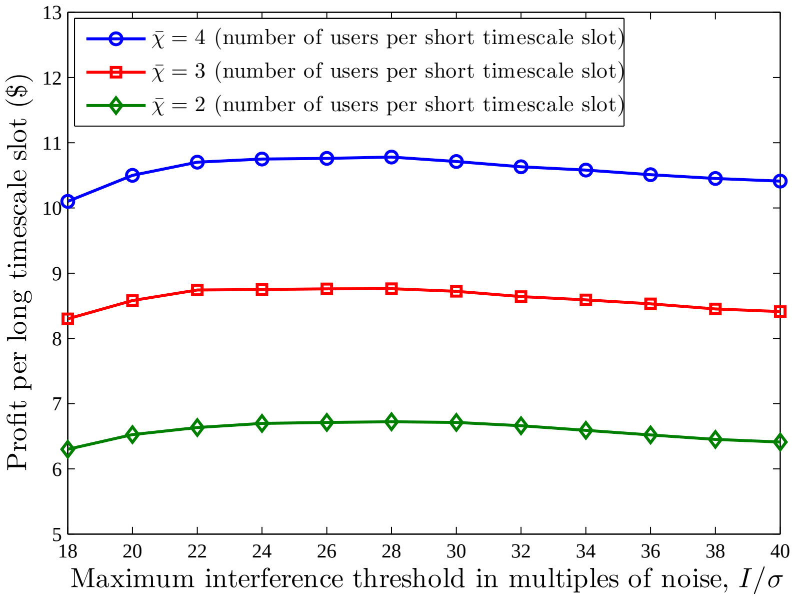

Figure 3.6: Impact of maximum interference threshold on optimal profit, $r^{\text{req}} = 1.5$ Mb/s, $\bar{\varepsilon}^2 = 0.05$.

A. In this part, we focus on evaluating the impact of maximum interference threshold on the optimal profit. As shown in Fig. 3.6, We find that the optimal profit increases and then decreases slightly as I increases. Thus, a suitable I can be obtained by running offline simulations. We can also find that the changes of the optimal profit with different I/σ^2 is not obvious, so the optimal profit is not very sensitive to the choice of I/σ^2.

3.3.3 Profit Comparison

We evaluate the performance of the proposed scheme with four other baseline schemes. In baseline schemes I and II, we maximize the profit obtained by the tenant. But in baseline scheme I, we only consider the CSI uncertainty. In baseline scheme II, we only consider the user traffic variation. We use these two baseline schemes to characterize the impact of user traffic variation and CSI uncertainty on the optimal profit. In baseline schemes III and IV,

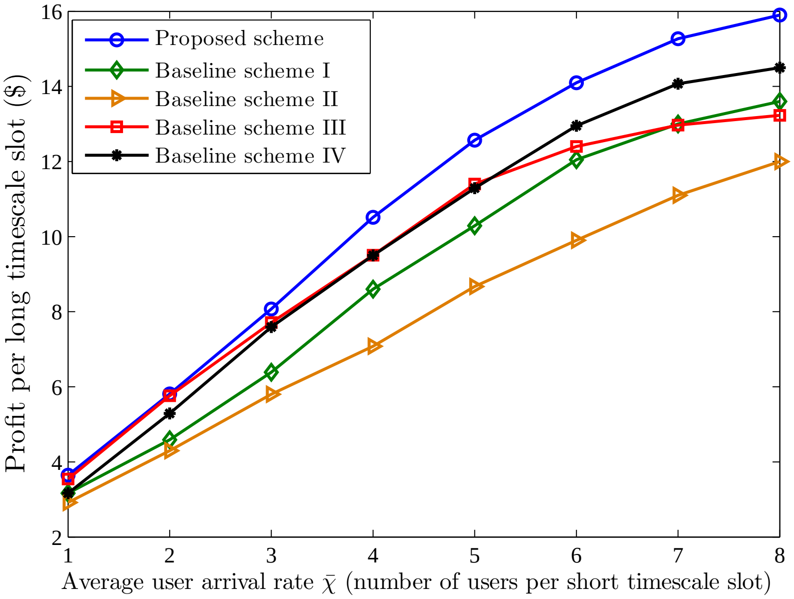

Figure 3.7: Profit comparison with different average user arrival rate $\bar{\chi}$ (number of users per short timescale slot), $r^{\text{req}} = 1.5$ Mb/s, $\bar{\varepsilon}^2 = 0.05$.

both CSI uncertainty and user traffic variation are considered, but the resource management schemes are slightly different from our proposed scheme. In baseline scheme III, we maximize the profit of the tenant while accepting all service requests from the users. This scheme is based on several resource management schemes for C-RAN which do not consider the user admission control [32, 33]. In baseline scheme IV, we maximize the profit of the tenant. In this scheme, the beamforming and user-centric RRH clustering are separated into two processes. User-centric RRH clustering is performed first for each new arrived user. Then, beamforming is designed for the user according to real-time network conditions.

Fig. 3.7 shows that the profit of the proposed scheme is larger than the profit of the four baseline schemes under different average user arrival rate $\bar{\chi}$ (number of arrived users per short timescale slot). Meanwhile, with the increasing of $\bar{\chi}$, the superiority of the proposed scheme

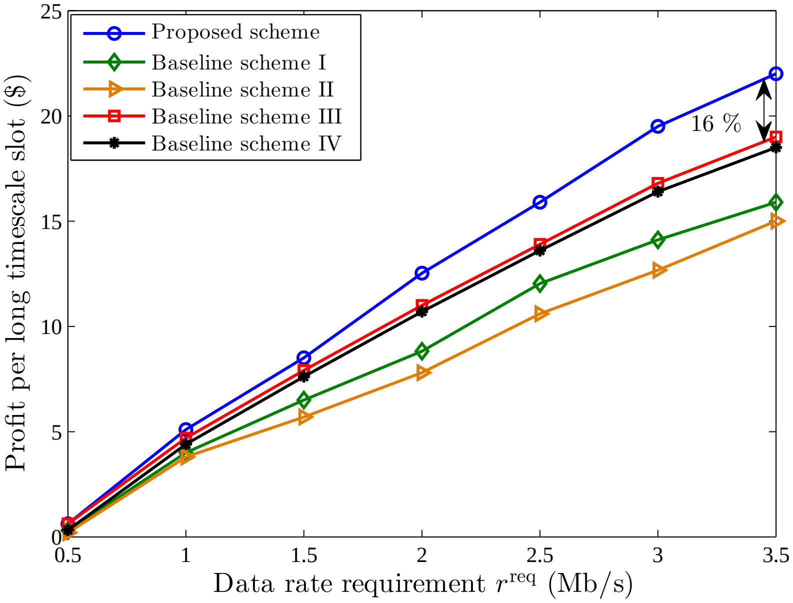

Figure 3.8: Profit comparison with different QoS requirements r^{req} (Mb/s), $\bar{\chi} = 3$ (number of users per short timescale slot), $\bar{\varepsilon}^2 = 0.05$.

in terms of the profit is more obvious compared with the four baseline schemes. It is because that higher revenue can be obtained from serving more users and the impact of the traffic variation and CSI uncertainty become more significant. Moreover, as the average user arrival rate increases, the increasing rate of the proposed scheme becomes slower. The reason for this behavior is that larger $\bar{\chi}$ leads to a larger number of users that may be close to each other in the coverage area. To mitigate interference, more resources need to be reserved, leading to higher resource reservation cost. We also find that the profit of baseline scheme III is close to the profit of the proposed scheme when $\bar{\chi}$ is no larger than 5. The reason is that when the number of users in the coverage area is small, the proposed scheme also tends to accept most of the users. The gap between the proposed scheme and baseline scheme III is due to the fact that the proposed scheme will reject users with really bad channel quality to save resources. The gap between

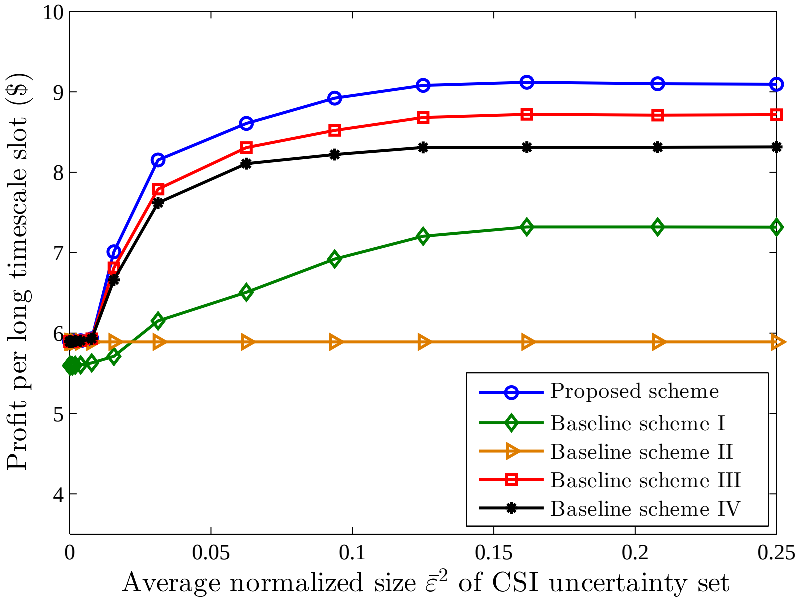

Figure 3.9: Profit comparison with different average normalized size $\bar{\varepsilon}^2$ of CSI uncertainty set, $r^{\text{req}} = 1.5$ Mb/s, $\bar{\chi} = 3$ (number of users per short timescale slot).

the proposed scheme and baseline scheme IV is due to the fact that our proposed scheme is more flexible to the network condition variations by designing user-centric RRH clustering and beamforming simultaneously.

Fig. 3.8 shows that the profit of the proposed scheme is larger than the profits of the four baseline schemes under different data rate requirement r^{req}. When r^{req} is large, the proposed scheme can achieve more than 16% profit improvement compared with the performance of the four baseline schemes. It is because higher data rate leads to higher revenue per user, making it more important to consider the user traffic variation and CSI uncertainty to obtain higher revenue from all users.

Fig. 3.9 shows that the proposed scheme achieves a higher profit compared with four baseline schemes under different choices of average normalized size $\bar{\varepsilon}^2$ of CSI uncertainty set.

Meanwhile, for most choices of $\bar{\varepsilon}^2$, the proposed scheme can achieve a higher profit compared with baseline scheme II, which does not consider the CSI uncertainty. When $\bar{\varepsilon}^2$ is close to zero, the CSI uncertainty is not fully considered for QoS guarantee in the proposed scheme. Thus, the gap of the profits between these two schemes is close to zero. With the increase of $\bar{\varepsilon}^2$, higher profit can be obtained by the proposed scheme according to revenue function (3.4). However, when $\bar{\varepsilon}^2$ is larger than 0.15, the profit will not increase further. It is because when $\bar{\varepsilon}^2$ is large, most of the CSI variations are considered in the CSI uncertainty set, and it is unnecessary to further increase $\bar{\varepsilon}^2$.

3.3.4 Resource Reservation and Allocation Performance

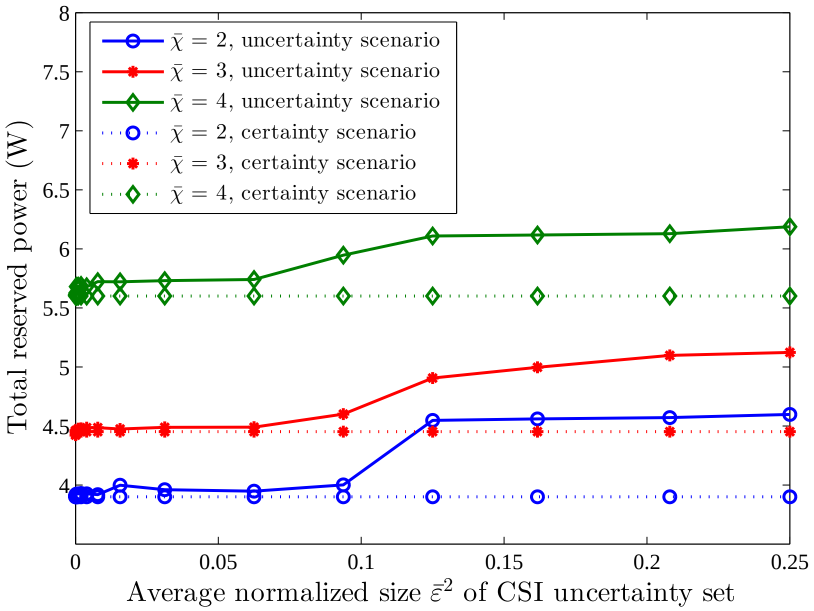

Figure 3.10: Total reserved power with different average user arrival rate $\bar{\chi}$ (number of users per short timescale slot) and different average normalized size $\bar{\varepsilon}^2$ of CSI uncertainty set, $r^{\text{req}} = 1.5$ Mb/s.

In this section, we evaluate the performance of the proposed scheme in terms of the resource

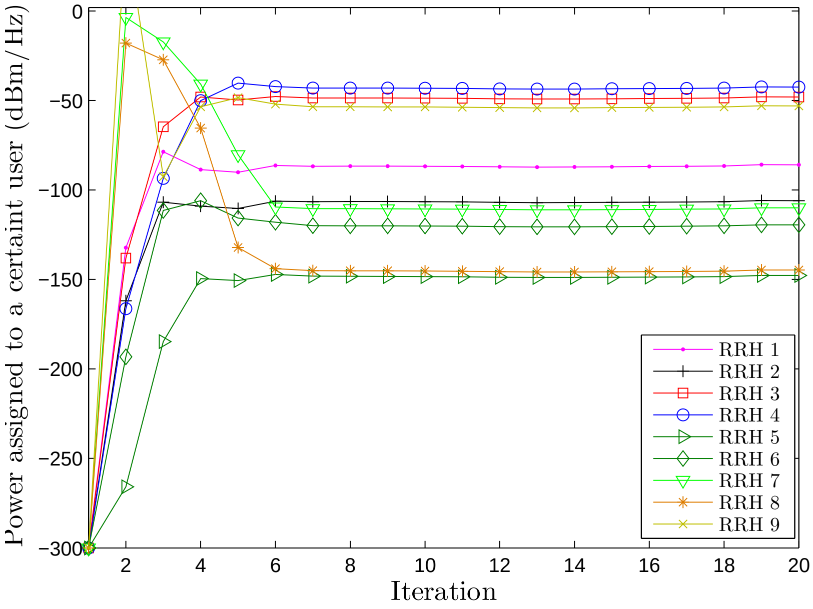

Figure 3.11: Power assigned to a user from all RRHs, $r^{\text{req}} = 1.5$ Mb/s, $\bar{\chi} = 3$ (number of users per short timescale slot), $\bar{\varepsilon}^2 = 0.05$.

reservation and allocation corresponding to different network conditions. Fig. 3.10 shows the decision of power reservation under different conditions of user traffic and CSI uncertainty. The total amount of reserved power increases as the average user arrival rate $\bar{\chi}$ increases in order to guarantee QoS requirements of more users. Meanwhile, the amount of reserved power increases as the average normalized size $\bar{\varepsilon}^2$ of CSI uncertainty set increases. When $\bar{\varepsilon}^2$ is getting smaller, the reserved amount of power will converge to the value of the reserved amount of power under the CSI certainty scenario. When $\bar{\varepsilon}^2$ is large, the increasing rate of the reserved amount of power will become small to avoid high power reservation cost.

Fig. 3.11 shows the power allocated to a certain user in a certain short timescale slot from all the RRHs by intra-slice resource allocation. From this figure, we notice that the power allocated to the user varies with different RRHs. The RRHs that are close to the user will allocate more power to the user and the RRHs that are far away from the user will almost allocate no power to the user to save energy and reduce resource reservation cost. Therefore, the beamforming designed in our proposed scheme can help achieve user-centric RRH clustering.

Chapter 4

Conclusion and Future Work

In this chapter, we conclude the thesis by summarizing the research work. We also suggest some possible extensions for future wok.

4.1 Conclusion

In this thesis, we proposed a resource management framework for network slicing in C-RAN from the perspective of each tenant. We maximized the profit of each tenant while characterizing the user traffic variation and CSI uncertainty by proposing a two-timescale resource management approach. To tackle two key challenges of user traffic variation and CSI uncertainty, we designed long timescale resource reservation with the statistical knowledge of user traffic, and designed short timescale intra-slice resource allocation, which is adaptive to arbitrary arrival/departure of users. The uncertainty set was applied for CSI uncertainty characterization to achieve robust intra-slice resource allocation for QoS guarantee. We formulated the profit maximization as a two-stage stochastic programming problem. We transformed the stochastic programming problem to a deterministic optimization problem, and performed QoS constraint approximation and semidefinite relaxation. We then applied branch-and-bound and primal-relaxed dual techniques to solve the problem. We conducted extensive simulations to evaluate the performance of our scheme. Based on the simulation results, we concluded that the convergence of the proposed scheme can be fast. The proposed scheme can achieve higher profit compared with other baseline schemes.

4.2 Future Work

In terms of future work, possible extensions are as follows:

1. In this thesis, we considered the required data rate as the QoS requirement needed to be guaranteed by each tenant. We proposed a profit model based on this QoS requirement. In 5G wireless systems, there are various types of services requiring different QoS. For example, ultra-reliable low-latency communications (URLLC) require low data transmission delay and high probability of successful data transmission. Massive machine type communications (mMTC) require efficient connectivity for a massive number of devices. Therefore, in our future work, we can consider new metrics for designing new QoS models under different types of services. Meanwhile, we can consider different profit models with revenue and cost functions according to different types of services.

2. In this thesis, we considered the profit maximization for a single tenant, and assumed that the resource reservation requests from the tenant can always be satisfied. In the future work, we can consider the scenario where multiple tenants coexist in the network. In this scenario, due to the limitation of network resources, the resource reservation request from each tenant may not always be satisfied. Therefore, we can apply auction and bidding techniques to formulate a new resource management framework for multiple tenants.

3. In this thesis, we assume constant durations of each long timescale slot and each short timescale slot. In the future work, we can consider the durations of the long timescale slot and short timescale slot as two variables to be optimized for achieving real-time adaptation to network conditions while controlling the computation cost.

Appendix A

Proof of Theorem 3.1

We follow approaches introduced in [45,46] to prove Theorem 3.1. For the optimal solution of problem (3.21), if the rank of any optimal beamforming matrix is larger than one, we consider the following optimization problem:

$$\begin{aligned} \underset{\mathbf{o}_k}{\text{minimize}} \quad & c_1 n_k + \frac{c_2}{LT} \sum_{l\in\mathcal{L}} \sum_{t\in\mathcal{T}_k} \sum_{u\in\mathcal{U}_{t,l}} \mathrm{Tr}(\mathbf{V}_{t,l,u}) - \sum_{l=1}^{L} \omega_l \sum_{t\in\mathcal{T}_k} \sum_{u\in\mathcal{U}_{t,l}} Y_{t,l,u}^{\text{new}}(a_{t,l,u}) \\ \text{subject to} \quad & \text{constraints (3.6c), (3.12d)}-\text{(3.12g), (3.17), (3.19),} \\ & \qquad\qquad \text{(3.20a)}-\text{(3.20e).} \end{aligned} \tag{A.1}$$

Problem (A.1) is a transformation of problem (3.21) by replacing $\sum_{b\in\mathcal{B}} c_2 p_{k,b}$ with $\frac{c_2}{LT}\sum_{l\in\mathcal{L}} \sum_{t\in\mathcal{T}_k} \sum_{u\in\mathcal{U}_{t,l}} \mathrm{Tr}(\mathbf{V}_{t,l,u})$ in the objective function. In problem (3.21), $\sum_{b\in\mathcal{B}} c_2 p_{k,b}$ in the objective function represents the cost of power reservation, where $\sum_{b\in\mathcal{B}} p_{k,b}$ serves as the upper bound for intra-slice power allocation in each short timescale slot for each realization of $\mathcal{U}_k^{\text{seq}}$. In problem (A.1), $\frac{1}{LT}\sum_{l\in\mathcal{L}} \sum_{t\in\mathcal{T}_k} \sum_{u\in\mathcal{U}_{t,l}} \mathrm{Tr}(\mathbf{V}_{t,l,u})$ represents the average power consumption over short timescale slots, and $\frac{c_2}{LT}\sum_{l\in\mathcal{L}}\sum_{t\in\mathcal{T}_k}\sum_{u\in\mathcal{U}_{t,l}}\mathrm{Tr}(\mathbf{V}_{t,l,u})$ can be regarded as the corresponding cost. We now prove that by solving problem (A.1), we obtain the optimal beamforming matrix, the rank of which is guaranteed to be either one or zero. Rank one solution indicates that the service request of the corresponding user is accepted, i.e., $a_{t,l,u} = 1$. Rank zero solution indicates that the service request of the corresponding user is rejected, i.e.,

$a_{t,l,u} = 0$.

We first obtain the Lagrangian of problem (A.1) but only focus on the terms that are related with out proof. The Lagrangian is given as follows:

$$\begin{aligned}
&\Gamma(\mathbf{V}_k^{\text{seq}}, \boldsymbol{\varphi}_k^{\text{seq}}, \boldsymbol{\upsilon}_k^{\text{seq}}, \boldsymbol{\xi}_k^{\text{seq}}, \mathbf{a}_k^{\text{seq}}, n_k, \mathbf{p}_k, \boldsymbol{\Lambda}) \\
&= \sum_{l\in\mathcal{L}} \sum_{t\in\mathcal{T}_k} \sum_{u\in\mathcal{U}_{t,l}} \mathbf{Tr}\left(\mathbf{V}_{t,l,u}\left(\frac{c_2}{LT}\mathbf{I}_{AB} + n_k \sum_{b\in\mathcal{B}} \varrho_{t,l,b}\mathbf{B}_b^{\text{H}}\mathbf{B}_b - \mathbf{Q}_{t,l,u}\mathbf{L}_{1,t,l,u}\mathbf{Q}_{t,l,u}^{\text{H}}\right.\right. \\
&\quad \left.\left. + \sum_{u'\in\mathcal{U}_{t,l}\setminus\{u\}} \mathbf{Q}_{t,l,u'}\mathbf{L}_{2,t,l,u'}\mathbf{Q}_{t,l,u'}^{\text{H}} - \mathbf{L}_{3,t,l,u}\right)\right) + \triangle,
\end{aligned} \tag{A.2}$$

where $\triangle$ represents the collection of the terms in the Lagrangian that are not related with our proof, $\boldsymbol{\Lambda}$ contains all the dual variables, $\varrho_{t,l,b}$ is the dual variable for constraint (3.20b), $\mathbf{L}_{1,t,l,u}$ is the dual matrix for constraint (3.17), $\mathbf{L}_{2,t,l,u}$ is the dual matrix for constraint (3.19), $\mathbf{L}_{3,t,l,u}$ is the dual matrix for constraint (3.20e).

We focus on the following KKT conditions that are related with our proof:

$$\nabla_{\mathbf{V}_{t,l,u}}\Gamma(\mathbf{V}_k^{\text{seq}}, \boldsymbol{\varphi}_k^{\text{seq}}, \boldsymbol{\upsilon}_k^{\text{seq}}, \boldsymbol{\xi}_k^{\text{seq}}, \mathbf{a}_k^{\text{seq}}, n_k, \mathbf{p}_k, \boldsymbol{\Lambda})|_{\bar{\mathbf{o}}_k,\bar{\boldsymbol{\Lambda}}} = \mathbf{O}_{AB},\ u\in\mathcal{U}_{t,l}, t\in\mathcal{T}_k, l\in\mathcal{L} \tag{A.3}$$

$$\bar{\mathbf{L}}_{3,t,l,u}\bar{\mathbf{V}}_{t,l,u} = \mathbf{O}_{AB},\ \ u\in\mathcal{U}_{t,l}, t\in\mathcal{T}_k, l\in\mathcal{L} \tag{A.4}$$

$$\left(\mathbf{S}_{1,t,l,u}(\bar{\boldsymbol{\varphi}}_k^{\text{seq}}, \bar{\boldsymbol{\upsilon}}_k^{\text{seq}}) + \mathbf{Q}_{t,l,u}^{\text{H}}\bar{\mathbf{V}}_{t,l,u}\mathbf{Q}_{t,l,u}\right)\bar{\mathbf{L}}_{1,t,l,u} = \mathbf{O}_{AB+1},\ \ u\in\mathcal{U}_{t,l}, t\in\mathcal{T}_k, l\in\mathcal{L} \tag{A.5}$$

$$\bar{\mathbf{L}}_{2,t,l,u} \succeq \mathbf{0},\ \ u\in\mathcal{U}_{t,l}, t\in\mathcal{T}_k, l\in\mathcal{L} \tag{A.6}$$

$$\bar{\mathbf{V}}_{t,l,u} \succeq \mathbf{0},\ \ u\in\mathcal{U}_{t,l}, t\in\mathcal{T}_k, l\in\mathcal{L} \tag{A.7}$$

$$\bar{\varrho}_{t,l,u} \geq 0,\ \ u\in\mathcal{U}_{t,l}, t\in\mathcal{T}_k, l\in\mathcal{L} \tag{A.8}$$

$$\bar{\upsilon}_{t,l,u} \geq 0,\ \ u\in\mathcal{U}_{t,l}, t\in\mathcal{T}_k, l\in\mathcal{L}, \tag{A.9}$$

where $\bar{\mathbf{o}}_k \triangleq (\bar{\mathbf{V}}_k^{\text{seq}}, \bar{\boldsymbol{\varphi}}_k^{\text{seq}}, \bar{\boldsymbol{\upsilon}}_k^{\text{seq}}, \bar{\boldsymbol{\xi}}_k^{\text{seq}}, \bar{\mathbf{a}}_k^{\text{seq}}, \bar{n}_k, \bar{\mathbf{p}}_k)$ and $\bar{\boldsymbol{\Lambda}}$ represents the optimal solutions of problem (A.1) and the corresponding optimal dual variables, respectively. We also have

$$\mathbf{S}_{1,t,l,u}(\bar{\boldsymbol{\varphi}}_k^{\text{seq}}, \bar{\boldsymbol{\upsilon}}_k^{\text{seq}}) = \begin{bmatrix} \bar{\upsilon}_{t,l,u}\mathbf{I}_{AB} & \mathbf{0}_{AB} \\ \mathbf{0}_{AB}^{\text{H}} & -\bar{\varphi}_{t,l,u}(I+\sigma^2) - \bar{\upsilon}_{t,l,u}\varepsilon_{t,l,u}^2 \end{bmatrix}. \tag{A.10}$$

By considering (A.3) and (A.4), we have

$$\mathbf{X}_{t,l,u}\bar{\mathbf{V}}_{t,l,u} = \mathbf{Q}_{t,l,u}\bar{\mathbf{L}}_{1,t,l,u}\mathbf{Q}_{t,l,u}^{\text{H}}\bar{\mathbf{V}}_{t,l,u},\ u \in \mathcal{U}_{t,l}, t \in \mathcal{T}_k, l \in \mathcal{L}, \tag{A.11}$$

where $\mathbf{X}_{t,l,u} \triangleq \frac{c_2}{LT}\mathbf{I}_{AB} + \bar{n}_k \sum_{b\in\mathcal{B}} \bar{\varrho}_{t,l,b}\mathbf{B}_b^{\text{H}}\mathbf{B}_b + \sum_{u'\in\mathcal{U}_{t,l}\setminus\{u\}} \mathbf{Q}_{t,l,u'}\bar{\mathbf{L}}_{2,t,l,u'}\mathbf{Q}_{t,l,u'}^{\text{H}}$. Moreover, since $\bar{n}_k > 0$, $\bar{\varrho}_{t,l,b} \geq 0$ and $\mathbf{Q}_{t,l,u'}\bar{\mathbf{L}}_{2,t,l,u'}\mathbf{Q}_{t,l,u'}^{\text{H}} \succeq \mathbf{0}$, we have $\mathbf{X}_{t,l,u} \succ \mathbf{0}$. Then, we have

$$\begin{aligned} &\text{Rank}(\bar{\mathbf{V}}_{t,l,u}) \\ &= \text{Rank}(\mathbf{X}_{t,l,u}\bar{\mathbf{V}}_{t,l,u}) \\ &= \text{Rank}(\mathbf{Q}_{t,l,u}\bar{\mathbf{L}}_{1,t,l,u}\mathbf{Q}_{t,l,u}^{\text{H}}\bar{\mathbf{V}}_{t,l,u}) \\ &\leq \min\left\{\text{Rank}(\mathbf{Q}_{t,l,u}\bar{\mathbf{L}}_{1,t,l,u}\mathbf{Q}_{t,l,u}^{\text{H}}), \text{Rank}(\bar{\mathbf{V}}_{t,l,u})\right\},\ u \in \mathcal{U}_{t,l}, t \in \mathcal{T}_k, l \in \mathcal{L}. \end{aligned} \tag{A.12}$$

To evaluate $\text{Rank}(\mathbf{Q}_{t,l,u}\bar{\mathbf{L}}_{1,t,l,u}\mathbf{Q}_{t,l,u}^{\text{H}})$, we post-multiply $\mathbf{Q}_{t,l,u}^{\text{H}}$ to (A.5) and obtain

$$\begin{aligned} \mathbf{S}_{t,l,u}(\bar{\boldsymbol{\varphi}}_k^{\text{seq}}, \bar{\boldsymbol{\upsilon}}_k^{\text{seq}})\bar{\mathbf{L}}_{1,t,l,u}\mathbf{Q}_{t,l,u}^{\text{H}} + \mathbf{Q}_{t,l,u}^{\text{H}}\bar{\mathbf{V}}_{t,l,u}\mathbf{Q}_{t,l,u}\bar{\mathbf{L}}_{1,t,l,u}\mathbf{Q}_{t,l,u}^{\text{H}} = \mathbf{O}_{(AB+1)\times AB}, \\ u \in \mathcal{U}_{t,l}, t \in \mathcal{T}_k, l \in \mathcal{L}. \end{aligned} \tag{A.13}$$

We then pre-multiply the left-hand side of (A.13) by $[\mathbf{I}_{AB}\,\mathbf{0}_{AB}]$. Considering that $\mathbf{Q}_{t,l,u} = [\mathbf{I}_{AB}\,\bar{\mathbf{h}}_{t,l,u}]$, we have

$$\begin{aligned}
&[\mathbf{I}_{AB}\,\mathbf{0}_{AB}]\mathbf{S}_{t,l,u}(\bar{\boldsymbol{\varphi}}_k^{\text{seq}}, \bar{\boldsymbol{\upsilon}}_k^{\text{seq}})\bar{\mathbf{L}}_{1,t,l,u}\mathbf{Q}^{\text{H}}_{t,l,u} + [\mathbf{I}_{AB}\,\mathbf{0}_{AB}]\mathbf{Q}^{\text{H}}_{t,l,u}\bar{\mathbf{V}}_{t,l,u}\mathbf{Q}_{t,l,u}\bar{\mathbf{L}}_{1,t,l,u}\mathbf{Q}^{\text{H}}_{t,l,u} = \mathbf{O}_{AB}\\
&\Leftrightarrow \bar{\upsilon}_{t,l,u}[\mathbf{I}_{AB}\,\mathbf{0}_{AB}]\bar{\mathbf{L}}_{1,t,l,u}\mathbf{Q}^{\text{H}}_{t,l,u} + \mathbf{I}_{AB}\bar{\mathbf{V}}_{t,l,u}\mathbf{Q}_{t,l,u}\bar{\mathbf{L}}_{1,t,l,u}\mathbf{Q}^{\text{H}}_{t,l,u} = \mathbf{O}_{AB}\\
&\Leftrightarrow \bar{\upsilon}_{t,l,u}\mathbf{Q}_{t,l,u}\bar{\mathbf{L}}_{1,t,l,u}\mathbf{Q}^{\text{H}}_{t,l,u} + \bar{\mathbf{V}}_{t,l,u}\mathbf{Q}_{t,l,u}\bar{\mathbf{L}}_{1,t,l,u}\mathbf{Q}^{\text{H}}_{t,l,u} = \bar{\upsilon}_{t,l,u}[\mathbf{O}_{AB}\ \bar{\mathbf{h}}_{t,l,u}]\\
&\Leftrightarrow (\bar{\upsilon}_{t,l,u}\mathbf{I}_{AB} + \bar{\mathbf{V}}_{t,l,u})\mathbf{Q}_{t,l,u}\bar{\mathbf{L}}_{1,t,l,u}\mathbf{Q}^{\text{H}}_{t,l,u} = \bar{\upsilon}_{t,l,u}[\mathbf{O}_{AB}\ \bar{\mathbf{h}}_{t,l,u}],\ u \in \mathcal{U}_{t,l}, t \in \mathcal{T}_k, l \in \mathcal{L}.
\end{aligned} \tag{A.14}$$

Based on the derivation in (A.14), we now calculate the rank of matrix $\mathbf{Q}_{t,l,u}\bar{\mathbf{L}}_{1,t,l,u}\mathbf{Q}^{\text{H}}_{t,l,u}$. Without the loss of generality, we consider two cases of $\bar{\upsilon}_{t,l,u}$. The first case is that $\bar{\upsilon}_{t,l,u} \neq 0$, and the second case is that $\bar{\upsilon}_{t,l,u} = 0$.

Case 1: $\bar{\upsilon}_{t,l,u} \neq 0$. According to (A.7) and (A.9), we can claim that the inverse matrix of $\bar{\upsilon}_{t,l,u}\mathbf{I}_{AB} + \bar{\mathbf{V}}_{t,l,u}$ exists. Based on the derivation in (A.14), we have

$$\begin{aligned}
&\text{Rank}(\mathbf{Q}_{t,l,u}\bar{\mathbf{L}}_{1,t,l,u}\mathbf{Q}^{\text{H}}_{t,l,u})\\
&\quad = \text{Rank}(\bar{\upsilon}_{t,l,u}(\bar{\upsilon}_{t,l,u}\mathbf{I}_{AB} + \bar{\mathbf{V}}_{t,l,u})^{-1}[\mathbf{O}_{AB}\ \bar{\mathbf{h}}_{t,l,u}])\\
&\quad \leq \text{Rank}([\mathbf{O}_{AB}\ \bar{\mathbf{h}}_{t,l,u}]) = 1.
\end{aligned} \tag{A.15}$$

Then we have

$$\begin{aligned}
\text{Rank}(\bar{\mathbf{V}}_{t,l,u}) \leq \min\left\{\text{Rank}(\mathbf{Q}_{t,l,u}\bar{\mathbf{L}}_{1,t,l,u}\mathbf{Q}^{\text{H}}_{t,l,u}), \text{Rank}(\bar{\mathbf{V}}_{t,l,u})\right\} \leq 1,&\\
u \in \mathcal{U}_{t,l}, t \in \mathcal{T}_k, l \in \mathcal{L}.&
\end{aligned} \tag{A.16}$$

For user $u \in \mathcal{U}_{t,l}$, if the service request is rejected, i.e., $\bar{a}_{t,l,u} = 0$, according to constraints (3.12d), (3.17) and (3.19), it is possible that $\text{Rank}(\bar{\mathbf{V}}_{t,l,u}) = 0$, indicating that no power will be assigned to the user. If the service request of user $u \in \mathcal{U}_{t,l}$ is accepted, i.e., $\bar{a}_{t,l,u} = 1$, in order to guarantee QoS requirement (3.12d), Rank$(\bar{\mathbf{V}}_{t,l,u}) = 1$ is required.

Case 2: $\bar{v}_{t,l,u} = 0$. In this case, we have

$$\bar{\mathbf{V}}_{t,l,u}\mathbf{Q}_{t,l,u}\bar{\mathbf{L}}_{1,t,l,u}\mathbf{Q}^{\mathrm{H}}_{t,l,u} = \mathbf{O}_{AB}. \tag{A.17}$$

According to (A.11), we have

$$\bar{\mathbf{V}}_{t,l,u}\mathbf{X}_{t,l,u} = \mathbf{O}_{AB}. \tag{A.18}$$

Since that $\mathbf{X}_{t,l,u} \succ 0$, we have $\bar{\mathbf{V}}_{t,l,u} = \mathbf{O}_{AB}$. For user $u \in \mathcal{U}_{t,l}$, $l \in \mathcal{L}$, $t \in \mathcal{T}_k$ whose service request is accepted, i.e., $\bar{a}_{t,l,u} = 1$, $\bar{\mathbf{V}}_{t,l,u} = \mathbf{O}_{AB}$ contradicts the QoS constraint (3.12d). In this case, $\bar{v}_{t,l,u} = 0$ will not happen. For user $u \in \mathcal{U}_{t,l}$, $l \in \mathcal{L}$, $t \in \mathcal{T}_k$ whose service request is rejected, $\bar{\mathbf{V}}_{t,l,u} = \mathbf{O}_{AB}$ is reasonable as no power will be assigned to the user. In this case, $\bar{v}_{t,l,u} = 0$ will happen.

Thus, for both Case 1 and Case 2, we have Rank$(\bar{\mathbf{V}}_{t,l,u}) \leq 1$.

Bibliography

[1] V. W. S. Wong, R. Schober, D. W. K. Ng, and L. Wang, *Key Technologies for 5G Wireless Systems*. Cambridge University Press, 2017.

[2] 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.

[3] V. Nguyen, A. Brunstrom, K. Grinnemo, and J. Taheri, “SDN/NFV-based mobile packet core network architectures: A survey,” *IEEE Communications Surveys Tutorials*, vol. 19, no. 3, pp. 1567–1602, Third quarter 2017.

[4] N. McKeown, T. Anderson, H. Balakrishnan, G. Parulkar, L. Peterson, J. Rexford, S. Shenker, and J. Turner, “OpenFlow: Enabling innovation in campus networks,” *ACM SIGCOMM Computer Communication Review*, vol. 38, no. 2, pp. 69–74, Mar. 2008.

[5] *OpenFlow Switch Specification Version 1. 5. 1.* Open Networking Foundation, 2015.

[6] R. Mijumbi, J. Serrat, J. Gorricho, N. Bouten, F. D. Turck, and R. Boutaba, “Network function virtualization: State-of-the-art and research challenges,” *IEEE Communications Surveys & Tutorials*, vol. 18, no. 1, pp. 236–262, First quarter 2016.

[7] 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.

[8] B. Dai and W. Yu, “Energy efficiency of downlink transmission strategies for cloud radio access networks,” *IEEE Journal on Selected Areas in Communications*, vol. 34, no. 4, pp. 1037–1050, Apr. 2016.

[9] X. Foukas, G. Patounas, A. Elmokashfi, and M. K. Marina, “Network slicing in 5G: Survey and challenges,” *IEEE Communications Magazine*, vol. 55, no. 5, pp. 94–100, May 2017.

[10] V. Sciancalepore, K. Samdanis, X. Perez, D. Bega, M. Gramaglia, and A. Banchs, “Mobile traffic forecasting for maximizing 5G network slicing resource utilization,” in *Proc. of IEEE Conference on Computer Communications (INFOCOM)*, Atlanta, GA, May 2017.

[11] R. Sherwood, G. Gibb, K.-K. Yap, G. Appenzeller, M. Casado, N. McKeown, and G. Parulkar, “FlowVisor: A network virtualization layer,” *Technical Report OPENFLOW-TR-2009-1*, Oct. 2009.

[12] A. Baumgartner, T. Bauschert, A. M. C. A. Koster, and V. S. Reddy, “Optimisation models for robust and survivable network slice design: A comparative analysis,” in *Proc. of IEEE Global Communications Conference (GLOBECOM)*, Singapore, Dec. 2017.

[13] S. Paris, A. Destounis, L. Maggi, G. S. Paschos, and J. Leguay, “Controlling flow reconfigurations in SDN,” in *Proc. of IEEE International Conference on Computer Communications (INFOCOM)*, San Francisco, CA, Apr. 2016.

[14] F. D. Pellegrini, L. Maggi, A. Massaro, D. Saucez, J. Leguay, and E. Altman, “Blind, adaptive and robust flow segmentation in datacenters,” in *Proc. of IEEE Conference on Computer Communications (INFOCOM)*, Honolulu, HI, Apr. 2018.

[15] K. Ravi, M. Rajesh, Z. Honghai, and R. Sampath, “NVS: A substrate for virtualizing wireless resources in cellular networks,” *IEEE/ACM Transactions on Networking*, vol. 20, no. 5, pp. 1333–1346, Oct. 2012.

[16] A. Gudipati, D. Perry, L. E. Li, and S. Katti, “SoftRAN: Software defined radio access network,” in *Proc. of ACM SIGCOMM Workshop on Hot Topics in Software Defined Networking (HotSDN)*, Hong Kong, China, Aug. 2013.

[17] P. Caballero, A. Banchs, G. de Veciana, X. Costa-Perez, P. Caballero, A. Banchs, G. de Veciana, and X. Costa-Perez, “Multi-tenant radio access network slicing: Statistical multiplexing of spatial loads,” *IEEE/ACM Transactions on Networking*, vol. 25, no. 5, pp. 3044–3058, Oct. 2017.

[18] H. Zhang, N. Liu, X. Chu, K. Long, A. H. Aghvami, and V. C. M. Leung, “Network slicing based 5G and future mobile networks: Mobility, resource management, and challenges,” *IEEE Communications Magazine*, vol. 55, no. 8, pp. 138–145, Aug. 2017.

[19] H. Xiang, W. Zhou, M. Daneshmand, and M. Peng, “Network slicing in fog radio access networks: Issues and challenges,” *IEEE Communications Magazine*, vol. 55, no. 12, pp. 110–116, Dec. 2017.

[20] W. Chen, X. Xu, C. Yuan, J. Liu, and X. Tao, “Virtualized radio resource pre-allocation for QoS based resource efficiency in mobile networks,” in *Proc. of IEEE Global Communications Conference (GLOBECOM)*, Singapore, Dec. 2017.

[21] Q. Zheng, K. Zheng, H. Zhang, and V. C. M. Leung, “Delay-optimal virtualized radio resource scheduling in software-defined vehicular networks via stochastic learning,” *IEEE Transactions on Vehicular Technology*, vol. 65, no. 10, pp. 7857–7867, Oct. 2016.

[22] P. Caballero, A. Banchs, G. de Veciana, and X. Costa-Prez, “Network slicing games: Enabling customization in multi-tenant networks,” in *Proc. of IEEE Conference on Computer Communications (INFOCOM)*, Atlanta, GA, May 2017.

[23] K. Zhu and E. Hossain, “Virtualization of 5G cellular networks as a hierarchical combinatorial auction,” *IEEE Transactions on Mobile Computing*, vol. 15, no. 10, pp. 2640–2654, Oct. 2016.

[24] Y. Zhang, S. Bi, and Y. J. A. Zhang, “Joint spectrum reservation and on-demand request for mobile virtual network operators,” *IEEE Transactions on Communications*, vol. 66, no. 7, pp. 2966–2977, Jul. 2018.

[25] W. Chen, X. Xu, C. Yuan, J. Liu, and X. Tao, "Virtualized radio resource pre-allocation for QoS based resource efficiency in mobile networks," in *Proc. of IEEE Global Communications Conference (GLOBECOM)*, Singapore, Dec. 2017.

[26] J. R. Birge and F. Louveaux, *Introduction to Stochastic Programming*. New York, NY, USA: Springer-Verlag, 1997.

[27] D. Chandramouli and T. Sun, "System architecture for the 5G system," *3GPP Release Rel-15 TS 23.501*, 2018.

[28] J. Groenendijk, "Study on management and orchestration of network slicing for next generation network," *3GPP Release Rel-14 TR 28.801*, 2017.

[29] M. Richart, J. Baliosian, J. Serrat, and J. Gorricho, "Resource slicing in virtual wireless networks: A survey," *IEEE Transactions on Network and Service Management*, vol. 13, no. 3, pp. 462–476, Sept. 2016.

[30] Y. Shi, J. Zhang, and K. B. Letaief, "Robust group sparse beamforming for multicast green cloud-RAN with imperfect CSI," *IEEE Transactions on Signal Processing*, vol. 63, no. 17, pp. 4647–4659, Sept. 2015.

[31] A. Liu and V. K. N. Lau, "Two-timescale user-centric RRH clustering and precoding optimization for cloud RAN via local stochastic cutting plane," *IEEE Transactions on Signal Processing*, vol. 66, no. 1, pp. 64–76, Jan. 2018.

[32] B. Dai and W. Yu, “Sparse beamforming and user-centric clustering for downlink cloud radio access network,” *IEEE Access*, vol. 2, pp. 1326–1339, Oct. 2014.

[33] Z. Wang, D. W. K. Ng, V. W. S. Wong, and R. Schober, “Robust beamforming design in C-RAN with sigmoidal utility and capacity-limited backhaul,” *IEEE Transactions on Wireless Communications*, vol. 16, no. 9, pp. 5583–5598, Sept. 2017.

[34] Y. L. Lee, J. Loo, T. C. Chuah, and L. C. Wang, “Dynamic network slicing for multitenant heterogeneous cloud radio access networks,” *IEEE Transactions on Wireless Communications*, vol. 17, no. 4, pp. 2146–2161, Jan. 2018.

[35] S. Costanzo, I. Fajjari, N. Aitsaadi, and R. Langar, “A network slicing prototype for a flexible cloud radio access network,” in *Proc. of IEEE Annual Consumer Communications Networking Conference (CCNC)*, Las Vegas, NV, Jan. 2018.

[36] M. Ezzaouia, C. Gueguen, M. E. Helou, M. Ammar, X. Lagrange, and A. Bouallegue, “A dynamic transmission strategy based on network slicing for cloud radio access networks,” in *Proc. of Wireless Days (WD)*, Dubai, UAE, Apr. 2018.

[37] B. Niu, Y. Zhou, H. Shah-Mansouri, and V. W. S. Wong, “A dynamic resource sharing mechanism for cloud radio access networks,” *IEEE Transactions on Wireless Communications*, vol. 15, no. 12, pp. 8325–8338, Dec. 2016.

[38] G. Gao, H. Hu, Y. Wen, and C. Westphal, “Resource provisioning and profit maximization for transcoding in clouds: A two-timescale approach,” *IEEE Transactions on Multimedia*, vol. 19, no. 4, pp. 836–848, Apr. 2017.

[39] J. Gong, J. S. Thompson, S. Zhou, and Z. Niu, "Base station sleeping and resource allocation in renewable energy powered cellular networks," *IEEE Transactions on Communications*, vol. 62, no. 11, pp. 3801–3813, Nov. 2014.

[40] A. Shapiro, D. Dentcheva, and A. Ruszczynski, *Lectures on Stochastic Programming: Modeling and Theory, Second Edition*. Philadelphia, PA, USA: Society for Industrial and Applied Mathematics, 2014.

[41] E. L. Lawler and D. E. Wood, "Branch-and-bound methods: A survey," *Operations Research*, vol. 14, pp. 699–719, Aug. 1966.

[42] I. Pólik and T. Terlaky, "A survey of the S-Lemma," *SIAM Review*, vol. 49, no. 3, pp. 371–418, Jul. 2007.

[43] Z. Luo, W. Ma, A. M. So, Y. Ye, and S. Zhang, "Semidefinite relaxation of quadratic optimization problems," *IEEE Signal Processing Magazine*, vol. 27, no. 3, pp. 20–34, May 2010.

[44] C. A. Floudas and V. Visweswaran, "Primal-relaxed dual global optimization approach," *Journal of Optimization Theory and Applications*, vol. 78, no. 2, pp. 187–225, Aug. 1993.

[45] D. W. K. Ng, E. S. Lo, and R. Schober, "Robust beamforming for secure communication in systems with wireless information and power transfer," *IEEE Transactions on Wireless Communications*, vol. 13, no. 8, pp. 4599–4615, Aug. 2014.

[46] E. Boshkovska, A. Koelpin, D. W. K. Ng, N. Zlatanov, and R. Schober, "Robust beamforming for SWIPT systems with non-linear energy harvesting model," in *Proc. of IEEE International Workshop on Signal Processing Advances in Wireless Communications (SPAWC)*, Edinburgh, U.K., Jul. 2016.