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

He Zhang and Vincent W.S. Wong, *Fellow, IEEE*

***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 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.**

***Index Terms*—C-RAN, network slicing, two timescales, stochastic programming, profit maximization.**

I. Introduction

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. This process is enabled by software-defined networking (SDN) and network function virtualization (NFV) [3], [4]. 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.

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 [5]. 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, located at the infrastructure provider, dynamically performs inter-slice resource virtualization and assigns the physical resources to each slice [6]. Then, the tenant performs intra-slice resource allocation among its subscribed users. 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.

Many works have been conducted on resource management for core network slicing [7]–[9]. The standardization of core network slicing has also been conducted within the 3rd Generation Partnership Project (3GPP) [10]–[12]. Several key concepts, such as network slice, network slice instance, and life-cycle management of network slice instance, are specified. Compared with core network slicing, RAN slicing is faced with new challenges due to time-varying channel conditions, user mobility, and interference. Conventional

Manuscript received on May 5, 2019; revised on Sept. 23, 2019 and Feb. 5, 2020; accepted on Mar. 22, 2020. This work was supported by the Natural Sciences and Engineering Research Council (NSERC) of Canada. The editor coordinating the review of this paper and approving it for publication was Jong-Moon Chung.

The authors are with the Department of Electrical and Computer Engineering, The University of British Columbia, Vancouver, Canada (email: {hezhang, vincentw}@ece.ubc.ca).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

approaches mainly consider inter-slice resource virtualization among tenants from the perspective of a centralized controller to achieve fairness among tenants [13]–[18]. For example, Gudipati *et al.* in [13] 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. Caballero *et al.* in [14] 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 [15] proposed a mobility management scheme and a joint power and sub-channel allocation scheme for RAN slicing to enhance resource efficiency. To achieve accurate resource demand estimation and efficient resource utilization, some studies have been conducted to design the resource reservation and intra-slice resource allocation from the perspective of each tenant [19], [20]. In [20], 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 [19] 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. 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 [21], [22]. Zhang *et al.* in [21] proposed a static spectrum reservation and dynamic resource requesting scheme for each tenant to maximize the aggregate utility of users. In [22], 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. However, these works neglect the characterization of the profit of each tenant and the impact of the uncertain and time-varying network conditions on the profit. To achieve profit maximization, each tenant should control the resource reservation cost and increase the revenue obtained from its subscribed users under the scenario of uncertain and time-varying network conditions.

Besides network slicing, cloud radio access network (C-RAN) is also a novel mobile network architecture for 5G wireless systems [23]. 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). 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, 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. By implementing multiple antennas at each RRH, the beamforming technique can be deployed to mitigate interference experienced by each user.

In this paper, 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.

The main contributions of this paper 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.

This paper is organized as follows. In Section II, we present the system model and the two-timescale framework, and

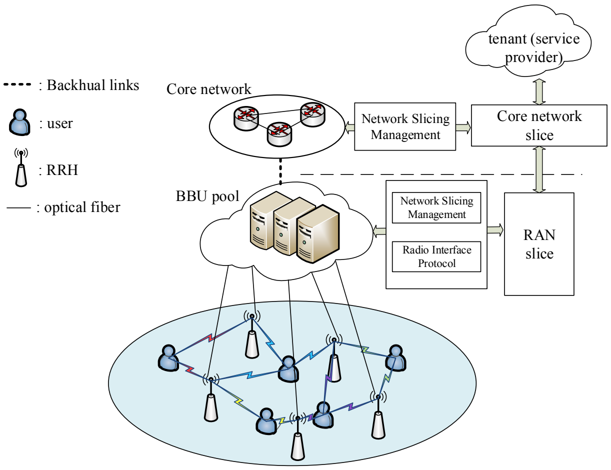


Fig. 1: Illustration of a CoMP-based C-RAN. Multiple RRHs cooperate together to transmit data to the users.

formulate the profit maximization problem. In Section III, we present the suboptimal solution of the problem. We evaluate the performance of the proposed scheme in Section IV, and conclude the paper in Section V.

Notations: The following notations are adopted: $\mathbf{X}^{\mathrm{H}}$, $\mathrm{Tr}(\mathbf{X})$, and $\mathrm{Rank}(\mathbf{X})$ represent the conjugate transpose, trace, and rank of matrix $\mathbf{X}$, respectively; $\mathbb{R}^{m\times n}$ is the set of m by n real matrices, $\mathbb{C}^{m\times n}$ is the set of m by n complex matrices; $\mathbb{H}^N$ denotes the set of N by N Hermitian matrices; $\mathbb{E}[\cdot]$ denotes the statistical expectation; $\mathbf{X} \succeq 0$ means that $\mathbf{X}$ is positive semidefinite. $\mathbf{I}_n$ is the $n \times n$ identity matrix, $\mathbf{0}_n$ denotes the $n \times 1$ all-zero vector; $\otimes$ stands for the Kronecker product; $\Re\{x\}$ stands for the real part of complex number x; $\mathcal{CN}(0, \sigma^2)$ is the zero-mean complex Gaussian distribution with variance σ^2.

II. System Model and Problem Formulation

A. Architecture of Network Slicing in C-RAN

We consider network slicing in a CoMP-based C-RAN system. As shown in Fig. 1, the system consists of a BBU pool, multiple RRHs, and a group of users. Multiple baseband signal processing modules are located at the BBU pool. The RRHs are connected to the BBU pool via optical fibers. The BBU pool connects to the core network via backhaul links. 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. In this 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. 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 the CoMP-based C-RAN. Each slice corresponds to a logical network with network resources and network functions allocated by an infrastructure provider. As shown in Fig. 1, each slice is owned and managed by a tenant (i.e., service provider) to support a specific type of service. In this paper, we consider the management of a RAN slice. According to 3GPP standards of 5G network slicing [10]–[12], the 5G management system, tied to the network functions virtualization management and orchestration (NFV-MANO) architectural framework, is implemented for network slices to achieve configuration, fault, lifecycle, and performance management. Each RAN slice consists of a management layer as well as layers of the radio interface protocol stack. The management layer is responsible of slice activation, configuration, and orchestration. The radio interface protocol stack consists of control plane functions of radio resource control (RRC), medium access control (MAC) layer, and physical (PHY) layer. In a CoMP-based C-RAN, since each RRH is only equipped with basic radio signal transmission and reception functions, the management system of the RAN slice is implemented in the cloud BBU pool, which is responsible of resource management including resource reservation and intra-slice resource allocation. In this paper, we assume that each tenant owns and manages a single slice. The proposed framework can be extended to the scenario where each tenant owns and manages multiple slices.

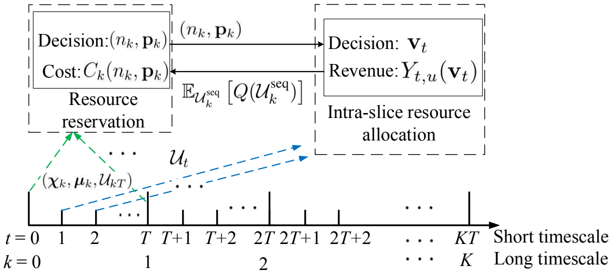

Decision: $(n_k, \mathbf{p}_k)$
Cost: $C_k(n_k, \mathbf{p}_k)$
Resource reservation
$(n_k, \mathbf{p}_k)$
$\mathbb{E}_{\mathcal{U}_k^{\mathrm{seq}}}\left[Q(\mathcal{U}_k^{\mathrm{seq}})\right]$
Decision: $\mathbf{v}_t$
Revenue: $Y_{t,u}(\mathbf{v}_t)$
Intra-slice resource allocation
$\mathcal{U}_t$
$(\boldsymbol{\chi}_k,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

Fig. 2: Two-timescale resource reservation and allocation.

We consider resource management for network slicing in C-RAN from the perspective of a single tenant and its single slice. 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 the channel conditions and QoS requirements.

B. Two-Timescale Framework

As shown in Fig. 2, the lifecycle of resource management for the considered slice is set to be one day. The reason that we choose the resource management lifecycle of a slice as one day is because this lifecycle exhibits the property of periodicity. The user traffic may vary during different time of the day. The variation of user traffic may follow a certain pattern during the day. We divide 24 hours into K long timescale slots (minutes). Each long timescale slot consists of T short timescale slots with the same duration. Resource reservation for the considered slice is updated at the beginning of each long timescale slot 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 can be captured so that the intra-slice resource allocation can be adaptive to the arbitrary user arrival and departure. In order to map the network slicing framework with the existing wireless systems, we use the transmission time interval (TTI) of 1 ms as the unit for a short timescale slot [24]. The duration of each short timescale slot can be set to any value between 1 second and 5 seconds (i.e., between 1000 and 5000 TTI units). Meanwhile, due to user mobility and fast channel fading, the uncertainty of CSI within each short timescale slot should also be considered. In this paper, 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. 2 will be given in the following part of this section.

1) User Traffic Model: In this paper, we consider the scenario where users arbitrarily arrive and leave the system [25], [26]. 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 [27]. Within the long timescale slot $k = 0, 1, \dots, K-1$, in region $m = 1, \dots, M$, 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. Since general distributions for both the arrival of users and sojourn time are assumed, we can design a resource management scheme that is applicable for different statistical models of user traffic.

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. By recording the service requests from the users and the service time for each user, the tenant can estimate the statistics of the user arrival process and sojourn time, as well as the user arrival rate $\chi_{k,m}$ and the mean sojourn time $\mu_{k,m}$.

2) Two-Timescale Resource Management: At the beginning of a long timescale slot k, the tenant obtains the knowledge of average 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. The precoded signal from RRHs to user u is given by $\mathbf{v}_{t,u} s_u = [\mathbf{v}_{t,u,1}^{\mathrm{H}} s_u \cdots \mathbf{v}_{t,u,B}^{\mathrm{H}} s_u]^{\mathrm{H}}$, where $s_u \in \mathbb{C}$ denotes the data symbol for user u. We assume that $\mathbb{E}[|s_u|^2] = 1$. Analog beamforming is used as the same data symbol s_u is fed to each antenna. We consider a single data stream for each user and each user is equipped with only one antenna. 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 signal received at user $u \in \mathcal{U}_t$ can be written as

$$\mathbf{h}_{t,u}^{\mathrm{H}} \mathbf{v}_{t,u} s_u + \sum_{u' \in \mathcal{U}_t \setminus \{u\}} \mathbf{h}_{t,u}^{\mathrm{H}} \mathbf{v}_{t,u'} s_u + n_u, \tag{1}$$

where the first term represents the desired signal and the second term represents the interfering signal, $n_u \sim \mathcal{CN}(0, \sigma^2)$ denotes the noise at user u with power σ^2. The data rate of user u in short timescale slot t can be obtained as follows [28]:

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

where the term $\sum_{u' \in \mathcal{U}_t \setminus \{u\}} |\mathbf{h}_{t,u}^{\mathrm{H}} \mathbf{v}_{t,u'}|^2$ represents the interference experienced by user u caused by data transmission from all the RRHs to other users. 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 paper, 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.

3) QoS Requirement under CSI Uncertainty: In this paper, 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

$$\mathcal{R}_{t,u} \triangleq \{\mathbf{h}_{t,u} \mid (\mathbf{h}_{t,u} - \bar{\mathbf{h}}_{t,u})^{\mathrm{H}}(\mathbf{h}_{t,u} - \bar{\mathbf{h}}_{t,u}) \leq \varepsilon_{t,u}^2\}, \quad (3)$$

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 (2) and (3), the QoS requirement under the CSI uncertainty can be modeled as

$$r_{t,u} \geq r^{\text{req}}, \quad \mathbf{h}_{t,u} \in \mathcal{R}_{t,u}. \quad (4)$$

Inequality (4) indicates that r^{req} should be satisfied for all the realizations of $\mathbf{h}_{t,u}$ in the set $\mathcal{R}_{t,u}$. By introducing the CSI uncertainty set, the QoS requirement can be depicted as deterministic constraint (4) without the necessity of knowing the statistical knowledge of the channel vector.

4) 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 [29]–[31]. In this section, we design a revenue model and a resource reservation cost model for the tenant.

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

$$Y_{t,u}(\mathbf{v}_t) = \begin{cases} p(\tilde{\varepsilon}_{t,u}) r^{\text{req}} \alpha, & r_{t,u} \geq r^{\text{req}},\ \mathbf{h}_{t,u} \in \mathcal{R}_{t,u} \\ -\beta, & \text{otherwise}, \end{cases} \quad (5)$$

where $\tilde{\varepsilon}_{t,u}^2 = \frac{\varepsilon_{t,u}^2}{\bar{\mathbf{h}}_{t,u}^{\mathrm{H}} \bar{\mathbf{h}}_{t,u}}$ 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 (5), 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 (4) 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

$$C_k(n_k, \mathbf{p}_k) = c_1 n_k + \sum_{b \in \mathcal{B}} c_2 p_{k,b}, \quad (6)$$

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.

C. 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:

$$\underset{n_k, \mathbf{p}_k}{\text{maximize}} \quad \mathbb{E}_{\mathcal{U}_k^{\text{seq}}}\left[Q(\mathcal{U}_k^{\text{seq}})\right] - C_k(n_k, \mathbf{p}_k) \quad (7a)$$

$$\text{subject to} \quad n_k \in \{0, \ldots, N\}, \quad (7b)$$

$$0 \leq p_{k,b} \leq P_b, \quad b \in \mathcal{B}, \quad (7c)$$

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:

$$\underset{\mathbf{v}_t, t \in \mathcal{T}_k}{\text{maximize}} \sum_{t \in \mathcal{T}_k} \sum_{u \in \mathcal{U}_t} Y_{t,u}(\mathbf{v}_t) \quad (8a)$$

$$\text{subject to } n_k \sum_{u \in \mathcal{U}_t} \mathrm{Tr}(\mathbf{v}_{t,u,b} \mathbf{v}_{t,u,b}^{\mathrm{H}}) \leq p_{k,b},\ b \in \mathcal{B}, t \in \mathcal{T}_k. \quad (8b)$$

Constraint (8b) represents the power constraint given the decisions of n_k and $\mathbf{p}_k$ made in the first stage.

By solving problem (7), the amount of reserved resources and the corresponding cost are determined, based on which second stage problem (8) 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.

III. Solution for the Profit Maximization Problem

A. Transformation into a Deterministic Problem

The two-stage stochastic programming problem cannot be solved directly due to the expectation of $Q(\mathcal{U}_k^{\text{seq}})$ in problem (7). 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 [32]. Based on the traffic model in Section II-B, 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:

$$\underset{n_k, \mathbf{p}_k, \mathbf{v}_k^{\text{seq}}}{\text{maximize}} \sum_{l=1}^{L} \omega_l \sum_{t \in \mathcal{T}_k} \sum_{u \in \mathcal{U}_{t,l}} Y_{t,u}(\mathbf{v}_{t,l}) - C_k(n_k, \mathbf{p}_k) \tag{9a}$$

$$\text{subject to} \quad n_k \sum_{u \in \mathcal{U}_{t,l}} \text{Tr}(\mathbf{v}_{t,l,u,b} \mathbf{v}_{t,l,u,b}^{\text{H}}) \leq p_{k,b}, \quad b \in \mathcal{B}, t \in \mathcal{T}_k, \, l \in \mathcal{L} \tag{9b}$$

$$\text{constraints (7b) and (7c),}$$

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

Problem (9) cannot be solved directly due to the nonconvexity of $Y_{t,u}(\mathbf{v}_{t,l})$. Based on the discussion in Section II-B4, the revenue function (5) 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 (4) is satisfied. The tenant will need to pay a penalty of β if constraint (4) 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 (5) is equivalent to

$$Y_{t,l,u}^{\text{new}}(a_{t,l,u}) = a_{t,l,u} p(\tilde{\varepsilon}_{t,l,u}) r^{\text{req}} \alpha - (1 - a_{t,l,u}) \beta, \tag{10}$$

with QoS constraint

$$n_k W \log \left(1 + \frac{|\mathbf{h}_{t,l,u}^{\text{H}} \mathbf{v}_{t,l,u}|^2}{\sum_{u' \in \mathcal{U}_{t,l} \setminus \{u\}} |\mathbf{h}_{t,l,u}^{\text{H}} \mathbf{v}_{t,l,u'}|^2 + \sigma^2}\right) \geq a_{t,l,u} r^{\text{req}}, \quad \mathbf{h}_{t,l,u} \in \mathcal{R}_{t,l,u}, u \in \mathcal{U}_{t,l}, t \in \mathcal{T}_k, l \in \mathcal{L}, \tag{11}$$

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. We reformulate problem (9) as follows:

$$\underset{\mathbf{a}_k^{\text{seq}}, n_k, \mathbf{p}_k, \mathbf{v}_k^{\text{seq}}}{\text{maximize}} \sum_{l=1}^{L} \omega_l \sum_{t \in \mathcal{T}_k} \sum_{u \in \mathcal{U}_{t,l}} Y_{t,l,u}^{\text{new}}(a_{t,l,u}) - C_k(n_k, \mathbf{p}_k) \tag{12a}$$

$$\text{subject to} \quad a_{t,l,u} \in \{0, 1\}, \;\; u \in \mathcal{U}_{t,l}, \, t \in \mathcal{T}_k, \, l \in \mathcal{L} \tag{12b}$$

$$\text{constraints (7b), (7c), (9b), (11),}$$

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 (12) is a mixed integer optimization problem due to integer variables $\mathbf{a}_k^{\text{seq}}$ and n_k. We use branch-and-bound technique [33] 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.

B. QoS Constraint Approximation and Semidefinite Relaxation

Based on the branch-and-bound technique, we focus on solving problem (12) with the relaxation of integer variables at each node. The relaxed optimization problem is still difficult to be solved as QoS constraint (11) is nonconvex. To tackle this challenge, we introduce a maximum interference threshold to achieve the QoS constraint approximation. The relaxed problem of (12) is formulated as follows:

$$\underset{\boldsymbol{\varphi}_k^{\text{seq}}, \mathbf{a}_k^{\text{seq}}, n_k, \mathbf{p}_k, \mathbf{v}_k^{\text{seq}}}{\text{maximize}} \sum_{l=1}^{L} \omega_l \sum_{t \in \mathcal{T}_k} \sum_{u \in \mathcal{U}_{t,l}} Y_{t,l,u}^{\text{new}}(a_{t,l,u}) - C_k(n_k, \mathbf{p}_k) \tag{13a}$$

$$\text{subject to} \quad \varphi_{t,l,u} \leq \frac{|\mathbf{h}_{t,l,u}^{\text{H}} \mathbf{v}_{t,l,u}|^2}{I + \sigma^2}, \quad \mathbf{h}_{t,l,u} \in \mathcal{R}_{t,l,u}, u \in \mathcal{U}_{t,l}, t \in \mathcal{T}_k, l \in \mathcal{L} \tag{13b}$$

$$\sum_{u' \in \mathcal{U}_{t,l} \setminus \{u\}} |\mathbf{h}_{t,l,u}^{\text{H}} \mathbf{v}_{t,l,u'}|^2 \leq I, \quad \mathbf{h}_{t,l,u} \in \mathcal{R}_{t,l,u}, u \in \mathcal{U}_{t,l}, t \in \mathcal{T}_k, l \in \mathcal{L} \tag{13c}$$

$$n_k W \log (1 + \varphi_{t,l,u}) \geq a_{t,l,u} r^{\text{req}}, \quad u \in \mathcal{U}_{t,l}, t \in \mathcal{T}_k, l \in \mathcal{L} \tag{13d}$$

$$0 \leq a_{t,l,u} \leq 1, u \in \mathcal{U}_{t,l}^{\text{relax}}, \, t \in \mathcal{T}_k, \, l \in \mathcal{L} \tag{13e}$$

$$a_{t,l,u} = d_{t,l,u}, \;\; u \in \mathcal{U}_{t,l} \backslash \mathcal{U}_{t,l}^{\text{relax}}, t \in \mathcal{T}_k, l \in \mathcal{L} \tag{13f}$$

$$0 \leq n_k \leq N, \tag{13g}$$

$$\text{constraints (7c) and (9b),}$$

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 (13) 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 (11) is relaxed as constraints (13b) (13c) and (13d) [34]. 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 (13b) and (13c) involves infinite number of constraints, making it difficult to directly solve problem (13). To tackle this problem, we apply

S-procedure [35] to transform constraints (13b) and (13c) into finite number of linear matrix inequality (LMI) constraints. The S-procedure is introduced in Lemma 1:

Lemma 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 $\mathbf{x} \in \mathbb{C}^{N\times 1}$:*

$$f_1(\mathbf{x}) = \mathbf{x}^H\mathbf{A}_1\mathbf{x} + 2\mathfrak{R}\{\mathbf{d}_1\mathbf{x}\} + y_1, \tag{14}$$

$$f_2(\mathbf{x}) = \mathbf{x}^H\mathbf{A}_2\mathbf{x} + 2\mathfrak{R}\{\mathbf{d}_2\mathbf{x}\} + y_2. \tag{15}$$

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

$$\theta \begin{bmatrix} \mathbf{A}_1 & \mathbf{d}_1 \\ \mathbf{d}_1^H & y_1 \end{bmatrix} - \begin{bmatrix} \mathbf{A}_2 & \mathbf{d}_2 \\ \mathbf{d}_2^H & y_2 \end{bmatrix} \succeq \mathbf{0}. \tag{16}$$

We denote that $\Delta\mathbf{h}_{t,l,u} = \mathbf{h}_{t,l,u} - \bar{\mathbf{h}}_{t,l,u}$. Then, by applying Lemma 1 to constraint (13b), we obtain the following implication:

$$\begin{aligned} &\Delta\mathbf{h}_{t,l,u}^{\mathrm{H}}\mathbf{I}_{AB}\Delta\mathbf{h}_{t,l,u} + 2\mathfrak{R}\{\mathbf{0}^{\mathrm{H}}\triangle\mathbf{h}_u\} - \varepsilon_{t,l,u}^2 \leq 0 \\ &\quad\Rightarrow -\Delta\mathbf{h}_{t,l,u}^{\mathrm{H}}\mathbf{V}_{t,l,u}\Delta\mathbf{h}_u - 2\mathfrak{R}\{(\mathbf{V}_{t,l,u}\bar{\mathbf{h}}_{t,l,u})^{\mathrm{H}}\Delta\mathbf{h}_u\} \\ &\qquad - \bar{\mathbf{h}}_{t,l,u}^{\mathrm{H}}\mathbf{V}_{t,l,u}\bar{\mathbf{h}}_{t,l,u} + \varphi_{t,l,u}(I+\sigma^2) \leq 0, \end{aligned} \tag{17}$$

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

$$\begin{aligned} &\begin{bmatrix} \upsilon_{t,l,u}\mathbf{I}_{AB} & \mathbf{0}_{AB} \\ \mathbf{0}_{AB}^{\mathrm{H}} & -\varphi_{t,l,u}(I+\sigma^2) - \upsilon_{t,l,u}\varepsilon_{t,l,u}^2 \end{bmatrix} \\ &\qquad + \mathbf{Q}_{t,l,u}^{\mathrm{H}}\mathbf{V}_{t,l,u}\mathbf{Q}_{t,l,u} \succeq \mathbf{0}, \quad u \in \mathcal{U}_{t,l}, t \in \mathcal{T}_k, l \in \mathcal{L}, \end{aligned} \tag{18}$$

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 1 to constraint (13c), we obtain the following implication:

$$\begin{aligned} &\Delta\mathbf{h}_{t,l,u}^{\mathrm{H}}\mathbf{I}_{AB}\Delta\mathbf{h}_{t,l,u} + 2\mathfrak{R}\{\mathbf{0}^{\mathrm{H}}\triangle\mathbf{h}_u\} - \varepsilon_{t,l,u}^2 \leq 0 \\ &\Rightarrow \Delta\mathbf{h}_{t,l,u}^{\mathrm{H}}\left(\textstyle\sum_{u'\in\mathcal{U}_{t,l}\setminus\{u\}}\mathbf{V}_{t,l,u'}\right)\Delta\mathbf{h}_{t,l,u} \\ &\quad +2\mathfrak{R}\left\{\left((\textstyle\sum_{u'\in\mathcal{U}_{t,l}\setminus\{u\}}\mathbf{V}_{t,l,u'})\bar{\mathbf{h}}_{t,l,u}\right)^{\mathrm{H}}\Delta\mathbf{h}_{t,l,u}\right\} \\ &\quad + \bar{\mathbf{h}}_{t,l,u}^{\mathrm{H}}\left(\textstyle\sum_{u'\in\mathcal{U}_{t,l}\setminus\{u\}}\mathbf{V}_{t,l,u'}\right)\bar{\mathbf{h}}_{t,l,u} - I \leq 0, \end{aligned} \tag{19}$$

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

$$\begin{aligned} &\begin{bmatrix} \xi_{t,l,u}\mathbf{I}_{AB} & \mathbf{0}_{AB} \\ \mathbf{0}_{AB}^{\mathrm{H}} & I - \xi_{t,l,u}\varepsilon_{t,l,u}^2 \end{bmatrix} \\ &\qquad -\mathbf{Q}_{t,l,u}^{\mathrm{H}}\left(\textstyle\sum_{u'\in\mathcal{U}_{t,l}\setminus\{u\}}\mathbf{V}_{t,l,u'}\right)\mathbf{Q}_{t,l,u} \succeq \mathbf{0}, \\ &\qquad u \in \mathcal{U}_{t,l}, t \in \mathcal{T}_k, l \in \mathcal{L}. \end{aligned} \tag{20}$$

Then, problem (13) is equivalent to

$$\underset{\mathbf{o}_k}{\text{minimize}}\ \ C_k(n_k,\mathbf{p}_k) - \sum_{l=1}^{L}\omega_l\sum_{t\in\mathcal{T}_k}\sum_{u\in\mathcal{U}_{t,l}}Y_{t,l,u}^{\text{new}}(a_{t,l,u}) \tag{21a}$$

$$\text{subject to}\ \ \text{constraints (7c), (13d)−(13g), (18), (20),}$$

$$n_k\sum_{u=1}^{|\mathcal{U}_{t,l}|}\mathrm{Tr}(\mathbf{B}_b^{\mathrm{H}}\mathbf{B}_b\mathbf{V}_{t,l,u}) \leq p_{k,b}, \qquad b\in\mathcal{B}, t\in\mathcal{T}_k, l\in\mathcal{L} \tag{21b}$$

$$\upsilon_{t,l,u} \geq 0, \qquad u \in \mathcal{U}_{t,l}, t \in \mathcal{T}_k, l \in \mathcal{L} \tag{21c}$$

$$\xi_{t,l,u} \geq 0, \qquad u \in \mathcal{U}_{t,l}, t \in \mathcal{T}_k, l \in \mathcal{L} \tag{21d}$$

$$\mathbf{V}_{t,l,u} \succeq \mathbf{0}, \qquad u \in \mathcal{U}_{t,l}, t \in \mathcal{T}_k, l \in \mathcal{L} \tag{21e}$$

$$\mathrm{Rank}(\mathbf{V}_{t,l,u}) \leq 1,\ \ u \in \mathcal{U}_{t,l}, t \in \mathcal{T}_k, l \in \mathcal{L}, \tag{21f}$$

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 $\mathrm{Tr}(\mathbf{v}_{t,l,u,b}\mathbf{v}_{t,l,u,b}^{\mathrm{H}}) = \mathrm{Tr}(\mathbf{B}_b^{\mathrm{H}}\mathbf{B}_b\mathbf{V}_{t,l,u})$. For constraint (21f), $\mathrm{Rank}(\mathbf{V}_{t,l,u}) = 0$ 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 (21) is still nonconvex due to constraint (21f). We adopt semidefinite relaxation (SDR) [36] by removing constraint (21f) to arrive at a tractable problem, given as:

$$\begin{aligned} &\underset{\mathbf{o}_k}{\text{minimize}}\ \ C_k(n_k,\mathbf{p}_k) - \sum_{l=1}^{L}\omega_l\sum_{t\in\mathcal{T}_k}\sum_{u\in\mathcal{U}_{t,l}}Y_{t,l,u}^{\text{new}}(a_{t,l,u}) \\ &\text{subject to}\ \ \text{constraints (7c), (13d)−(13g), (18), (20),} \\ &\qquad\qquad \text{(21b)−(21e).} \end{aligned} \tag{22}$$

For the optimal solution of problem (22), 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 (21) and (22) have the same optimal solution and the same optimal objective value. Otherwise, the optimal objective value of problem (21) serves as the lower bound of problem (22).

C. Primal-Relaxed Dual Technique

Problem (22) is still difficult to be solved directly due to the nonconvexity of constraints (13d) and (21b). One observation is that by fixing variables n_k and $\mathbf{p}_k$, problem (22) 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 (22) 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 [37]. The main idea of the primal-relaxed dual technique is to convert the original problem into primal and relaxed dual subproblems, which correspond to the upper and lower bounds, respectively, on the global optimal value. By iteratively solving the primal subproblem and the relaxed dual subproblem, the upper and lower bounds converge to the optimal value. Specifically, in each iteration, the primal subproblem is obtained by fixing variables $\boldsymbol{\varphi}_k^{\text{seq}}$, $\boldsymbol{\upsilon}_k^{\text{seq}}$, $\boldsymbol{\xi}_k^{\text{seq}}$, $\mathbf{a}_k^{\text{seq}}$, $\mathbf{V}_k^{\text{seq}}$ of problem (22) as constant values, which are obtained from the solution of the relaxed dual subproblem in the previous iteration. The primal problem is a linear programming problem with variables n_k and $\mathbf{p}_k$. The relaxed dual subproblem is obtained by first deriving the Lagrangian of the original problem, then fixing $\mathbf{p}_k$ as a constant value and choosing n_k between its valid upper and lower bounds. The value of $\mathbf{p}_k$ is obtained from the solution of the primal problem in the previous iteration.

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 (22) with respect to n_k and $\mathbf{p}_k$, which is formulated as follows:

$$\begin{aligned}&\underset{n_k,\mathbf{p}_k}{\text{minimize}} \;\; C_k(n_k,\mathbf{p}_k) - \sum_{l=1}^{L}\omega_l \sum_{t\in\mathcal{T}_k}\sum_{u\in\mathcal{U}_{t,l}} Y^{\text{new}}_{t,l,u}(a_{t,l,u})\\ &\text{subject to} \;\; \text{constraints (7c), (13d), (13g), (21b).}\end{aligned} \tag{23}$$

The obtained optimal value is denoted as f^{upper}, serving as an upper bound of problem (22). The corresponding solutions of Lagrange multipliers for constraints (13d), (21b), (13g), and (7c) 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 (22), we derive the Lagrangian of problem (22) with constraints (13d), (21b), (13g), and (7c), given as

$$\begin{aligned}&L(\boldsymbol{\varphi}_k^{\text{seq}},\mathbf{a}_k^{\text{seq}},\mathbf{V}_k^{\text{seq}},n_k,\mathbf{p}_k,\boldsymbol{\lambda})\\ &= n_k G_1(\boldsymbol{\varphi}_k^{\text{seq}},\mathbf{V}_k^{\text{seq}},\boldsymbol{\lambda}) + G_2(\mathbf{a}_k^{\text{seq}},\mathbf{p}_k,\boldsymbol{\lambda}),\end{aligned} \tag{24}$$

where

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

and

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

With the Lagrangian, we further have

$$\begin{aligned}&\inf_{0\le n_k\le N} L(\boldsymbol{\varphi}_k^{\text{seq}},\mathbf{a}_k^{\text{seq}},\mathbf{V}_k^{\text{seq}},n_k,\mathbf{p}_k,\boldsymbol{\lambda})\\ &= \inf_{0\le n_k\le N} n_k G_1(\boldsymbol{\varphi}_k^{\text{seq}},\mathbf{V}_k^{\text{seq}},\boldsymbol{\lambda}) + G_2(\mathbf{a}_k^{\text{seq}},\mathbf{p}_k,\boldsymbol{\lambda})\\ &= \frac{N-\delta N}{2}G_1(\boldsymbol{\varphi}_k^{\text{seq}},\mathbf{V}_k^{\text{seq}},\boldsymbol{\lambda}) + G_2(\mathbf{a}_k^{\text{seq}},\mathbf{p}_k,\boldsymbol{\lambda}),\end{aligned} \tag{25}$$

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 [37], we obtain the relaxed dual problem of problem (22) as follows:

$$\underset{\boldsymbol{\varphi}_k^{\text{seq}},\boldsymbol{\upsilon}_k^{\text{seq}},\boldsymbol{\xi}_k^{\text{seq}},\mathbf{a}_k^{\text{seq}},\mathbf{V}_k^{\text{seq}},\delta}{\text{minimize}} \;\; \frac{N-\delta N}{2}G_1(\boldsymbol{\varphi}_k^{\text{seq}},\mathbf{V}_k^{\text{seq}}) + G_2(\mathbf{a}_k^{\text{seq}}) \tag{26a}$$

$$\begin{aligned}\text{subject to} \quad &\text{constraints (13e), (13f), (18), (20),}\\ &\text{(21c)}-\text{(21e)}\end{aligned}$$

$$\delta G_1(\boldsymbol{\varphi}_k^{\text{seq}},\mathbf{V}_k^{\text{seq}}) \ge 0, \tag{26b}$$

$$\delta\in\{-1,1\}. \tag{26c}$$

The optimal value of problem (26) is denoted as f^{lower}, serving as a lower bound of problem (22). We iteratively solve the primal problem (23) and the relaxed dual problem (26) 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. 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 (7) and (8). We first transform the problem into deterministic optimization problem (9). Due to the nonconvexity of revenue function (5), then we transform the revenue function into a linear function with QoS constraint (11) by introducing a user admission control variable $a_{t,l,u}$ for each user. Moreover, problem (9) can be transformed into problem (12). Due to the nonconvexity of QoS constraint (11), 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 (13). 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 (13) for each node. To solve this problem, we apply S-procedure and SDR to obtain problem (22), solved by applying primal-relaxed dual technique. In Fig. 3, a bidirectional arrow represents a transformation into a equivalent problem. The unidirectional arrows from problem (12) to problem (13), and from problem (21) to problem (22), represent transformations involving approximations.

D. Joint Resource Reservation and Allocation Algorithm

In this section, we design an 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 an algorithm shown in Algorithm 1, which applies the primal-relaxed dual technique for each node. We then design the global algorithm for resource reservation and intra-slice resource allocation in long timescale slot $k = 0,\ldots,K$ based on the branch-and-bound technique and a call of Algorithm 1 in the inner iteration. This algorithm is shown in Algorithm 2.

In Algorithms 1 and 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 2, steps $7-14$ depict the process to calculate the optimal solutions for the two child nodes generated from the last

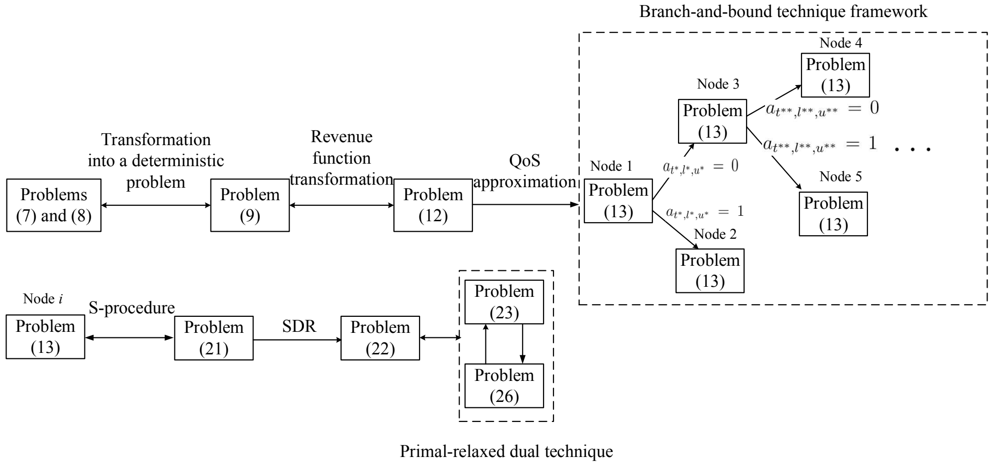


Fig. 3: The transformation and relaxation steps taken from problems (7) and (8) to obtain the solutions of the profit maximization problem.

Algorithm 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: Set $j := 1$.
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 (18), (20), (21c)−(21e); 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 (23) 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 (26) 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 while
11: Return $f^{\text{upper}}(i,j)$, $f^{\text{lower}}(i,j)$, and optimal solution $\mathbf{o}_k(i,j) := (\boldsymbol{\varphi}_k^{\text{seq}}(i,j)$,$\boldsymbol{\upsilon}_k^{\text{seq}}(i,j)$, $\boldsymbol{\xi}_k^{\text{seq}}(i,j)$, $\mathbf{V}_k^{\text{seq}}(i,j)$, $\mathbf{a}_k^{\text{seq}}(i,j)$, $n_k(i,j)$, $\mathbf{p}_k(i,j))$.
```

Algorithm 2 Global Algorithm for Resource Reservation and Intra-Slice Resource Allocation in Long Timescale Slot k

```
1: Set $i := 1$, in which $i$ represents the index of the node of the branch-and-bound technique.
2: 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 $a_{t,l,u}(i)$, $u \in \mathcal{U}_{t,l}$, $t \in \mathcal{T}_k$, $l \in \mathcal{L}$.
3: $\mathcal{U}_{t,l}^{\text{relax}}(i) := \mathcal{U}_{t,l}$, $t \in \mathcal{T}_k$, $l \in \mathcal{L}$.
4: Initialize $\mathcal{D}(i) := \emptyset$ to record the set of $(d_{t,l,u}, t, l, u)$ at the current node.
5: Initialize $\mathcal{F}_{\text{node}} := \emptyset$ to record the optimal values and solutions of ending nodes and the indexes of the nodes.
6: while $\exists\ a_{t,l,u}(i) \notin \{0,1\}$, $\forall u$, $t$, $l$, do
7:     $s := 1$.
8:     while $s \leq 2$ do
9:         Call Algorithm 1, with input $\mathcal{D}(i)$, $\mathbf{a}_k^{\text{seq}}(i)$, and $\mathcal{U}_{t,l}^{\text{relax}}(i)$, $t \in \mathcal{T}_k$, $l \in \mathcal{L}$.
10:        $f := \frac{f^{\text{upper}(i,j)} + f^{\text{lower}(i,j)}}{2}$.
11:        $\mathbf{o}_k := \mathbf{o}_k(i,j)$.
12:        $\mathcal{F}_{\text{node}} := \mathcal{F}_{\text{node}} \bigcup \{(f, \mathbf{o}_k, i)\}$.
13:        $i := i + 1$, $s := s + 1$.
14:    end while
15:    $(f^*, \mathbf{o}^*, i^*) := \arg\min_{f^*} \{\mathcal{F}_{\text{node}}\}$; 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 $\mathbf{o}^*$.
16:    $\mathcal{D}(i) := \mathcal{D}(i^*) \bigcup \{(0, t^*, l^*, u^*)\}$, $\mathcal{D}(i+1) := \mathcal{D}(i^*) \bigcup \{(1, t^*, l^*, u^*)\}$.
17:    $\mathcal{U}_{t^*,l^*}^{\text{relax}}(i) := \mathcal{U}_{t^*,l^*}^{\text{relax}}(i^*) \backslash \{u^*\}$, $\mathcal{U}_{t^*,l^*}^{\text{relax}}(i+1) := \mathcal{U}_{t^*,l^*}^{\text{relax}}(i^*) \backslash \{u^*\}$.
18:    $\mathcal{F}_{\text{node}} := \mathcal{F}_{\text{node}} \backslash \{(f^*, \mathbf{o}^*, i^*)\}$.
19: end while
20: Return $-f^*$.
```

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 2 is dominated by the branch-and-bound technique, and is in $\mathcal{O}(2^n)$, where n is the total number of user admission control variables. However, in practice, the algorithm can run fast as only a small number of nodes are searched before reaching the optimal solutions.

E. Discussion of Implementation Aspects

Based on the 3GPP standard on 5G network slicing [10]–[12] and the architecture of CoMP-based C-RAN, we now discuss the implementation aspects of our proposed framework. At the beginning of each short timescale slot, each user provides the network slice selection assistance information (NSSAI) parameters to the network to help it select the RAN slice. After the user has been granted access to the slice, the RRC function of the control plane implemented at the BBU pool starts to collect the CSI feedback sent by the user to each RRH, based on which the beamforming algorithm is invoked for the user at the BBU pool for resource allocation. The BBU pool then sends the user data messages and resource allocation decisions to multiple RRHs by fronthaul links, based on which RRHs cooperatively form the beamforming vector and transmit the downlink data messages to the users. In the meantime, the sojourn time and the number of users which accessed the slice are recorded in the management system for the estimation of user traffic statistics. The estimated information is then used for the resource reservation decision made at the beginning of each long timescale slot.

IV. Performance Evaluation

A. Simulation Environment

The coverage area of a C-RAN 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 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 $\bar{\varepsilon}^2_{t,l,u}$ 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. The simulations are performed using Matlab.

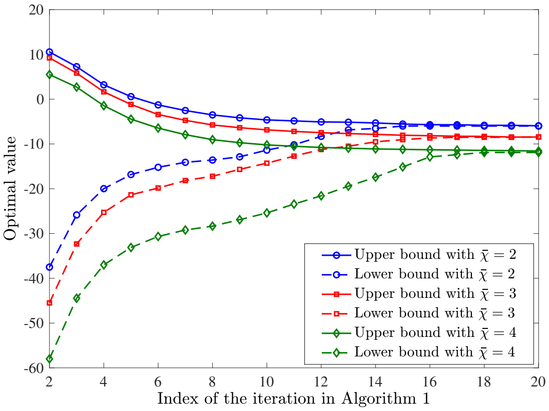


Fig. 4: Convergence of Algorithm 1 with different average user arrival rate $\bar{\chi}$ (number of users per short timescale slot), $r^{\text{req}}=1.5$ Mb/s, $\bar{\varepsilon}^2=0.05$.

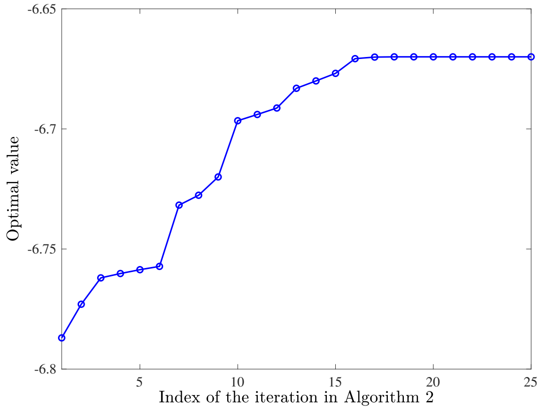


Fig. 5: Outer iteration convergence of Algorithm 2, $r^{\text{req}}=1.5$ Mb/s, $\bar{\varepsilon}^2=0.05$, $\bar{\chi}=2$ (number of users per short timescale slot).

B. Algorithm Properties

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 1 and 2. The simulation results are shown in Figs. 4, 5 and 6. Fig. 4 shows the convergence of Algorithm 1. Each iteration represents the process of solving problems (23) and (26) 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 area, thus a larger number of variables to be solved at each iteration in Algorithm 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 1.

Figs. 5 and 6 show the outer iteration convergence of Algorithm 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 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. 6, $\bar{\chi}=4$, which is larger than $\bar{\chi}=2$ in Fig. 5. So, there is a larger number of users in the system, which leads to a larger number 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. 6 is slower than that in Fig. 5. 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 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.

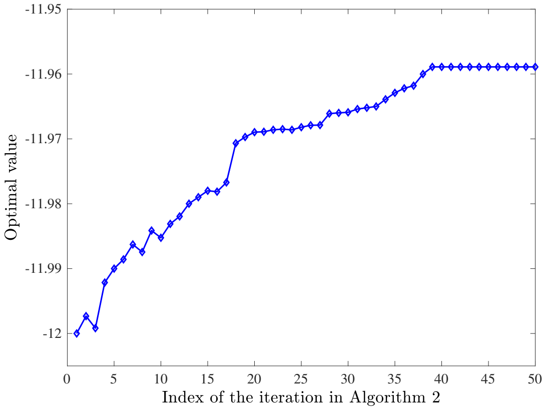


Fig. 6: Outer iteration convergence of Algorithm 2, $r^{\text{req}} = 1.5$ Mb/s, $\bar{\varepsilon}^2 = 0.05$, $\bar{\chi} = 4$ (number of users per short timescale slot).

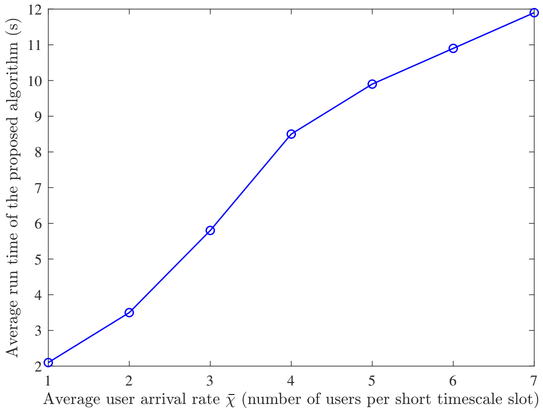


Fig. 7: Average run time of the proposed algorithm with different average user arrival rate $\bar{\chi}$ (number of users per short timescale slot), $r^{\text{req}} = 1.5$ Mb/s, $\bar{\varepsilon}^2 = 0.05$.

In order to show the time complexity of the proposed approach, results of the average run time or execution time of the proposed algorithm under different average user arrival rates are plotted in Fig. 7. Results show the average run time of the proposed algorithm increases linearly with the average user arrival rate. Compared with the duration of each long timescale slot, which is 20 minutes in our simulation, several seconds of run time of the proposed algorithm is acceptable for practical implementation.

As discussed in Section III-B, we cannot obtain the exact optimal solution and the optimal profit for two reasons. First, we introduced a maximum interference threshold. Second, we applied semidefinite relaxation. We now evaluate the impact of maximum interference threshold on the optimal profit. As shown in Fig. 8, 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.

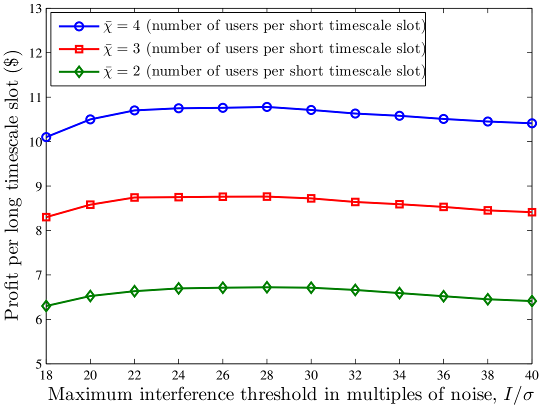


Fig. 8: Impact of maximum interference threshold on the profit, $r^{\text{req}} = 1.5$ Mb/s, $\bar{\varepsilon}^2 = 0.05$.

C. Profit Comparison

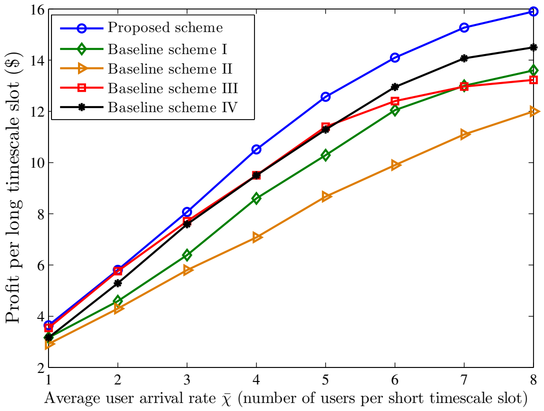


Fig. 9: Profit comparison with different average user arrival rate $\bar{\chi}$ (number of users per short timescale slot), $r^{\text{req}} = 1.5$ Mb/s, $\bar{\varepsilon}^2 = 0.05$.

We applied four baseline schemes to evaluate the performance of our proposed algorithm. Baseline schemes I and II are developed from our proposed scheme. They both maximize the profit of the considered tenant via long timescale resource reservation and short timescale intra-slice resource allocation. The profit models of these two schemes are similar with that of our proposed scheme. Instead of addressing the two issues of CSI uncertainty and user traffic variation, baseline scheme I only addresses the issue of CSI uncertainty and baseline scheme II only addresses the issue of user traffic variation. The intention of comparing our proposed scheme with baseline schemes I and II is to characterize the impact of CSI uncertainty and user traffic variation on the optimal solution separately. Baseline schemes III and IV maximize the profit of the considered tenant while addressing CSI uncertainty and user traffic variation. However, they use different intra-slice resource allocation schemes from [28] and [38]. In baseline scheme III, user admission control is not considered and the beamforming algorithm from [28] is invoked. In baseline scheme IV, user-centric RRH clustering is first performed, after which the beamforming scheme from [38] is invoked. Note that in our proposed scheme, by designing beamforming vectors, user-centric RRH clustering is included in the beamforming scheme.

Fig. 9 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 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 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. 10 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. 11 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 (5). 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 $\bar{\varepsilon}^2$.

D. Resource Reservation and Allocation Performance

In this subsection, we evaluate the performance of the proposed scheme in terms of the resource reservation and allocation corresponding to different network conditions. Fig. 12 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. 13 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

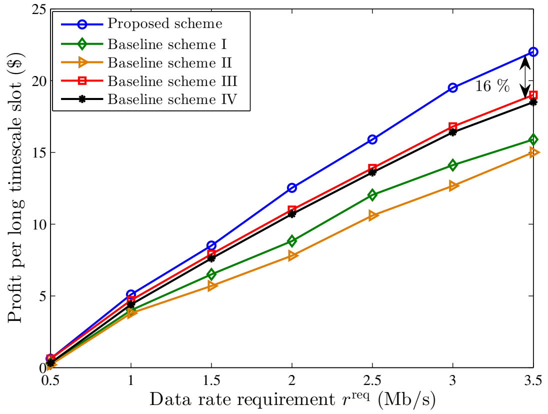


Fig. 10: Profit comparison with different QoS requirements r^{req} (Mb/s), $\bar{\chi} = 3$ (number of users per short timescale slot), $\bar{\varepsilon}^2 = 0.05$.

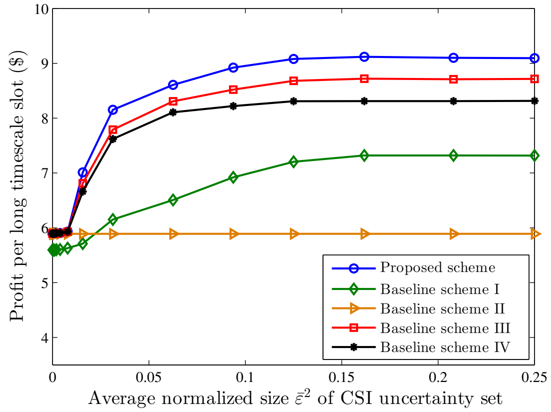


Fig. 11: Profit comparison with different average normalized size $\bar{\varepsilon}^2$ of CSI uncertainty set, $r^{\text{req}} = 1.5$ Mb/s, $\bar{\chi} = 3$ (number of users per short timescale slot).

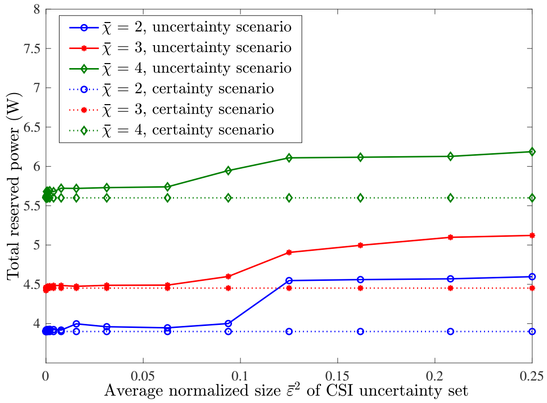


Fig. 12: Total reserved power with different average user arrival rate $\bar{\chi}$ (number of users per short timescale slot) and different average normalized size $\bar{\varepsilon}^2$ of CSI uncertainty set, $r^{\text{req}} = 1.5$ Mb/s.

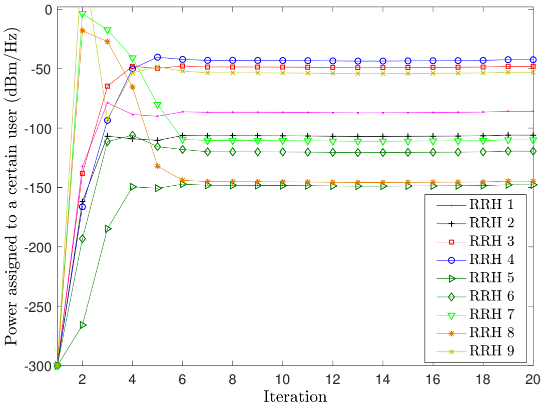


Fig. 13: Power assigned to a user from all RRHs, $r^{\text{req}} = 1.5$ Mb/s, $\bar{\chi} = 3$ (number of users per short timescale slot), $\bar{\varepsilon}^2 = 0.05$.

designed in our proposed scheme can help achieve user-centric RRH clustering.

Fig. 14 shows the sensitivity of the proposed scheme to different values of resource reservation costs and penalty of failing to serve a user in a long timescale slot. When the resource reservation costs c_1 and c_2 or the penalty β increase, the profit obtained by the tenant decreases. Compared with users who are closer to the RRHs, users who are far away from the RRHs may experience poor channel quality. In order to guarantee the required data rate for the users who are far away from the RRHs, more resources need to be allocated, leading to higher resource reservation cost and low profit. When the value of penalty is low, the tenant may decide to reject the service requests from the users who are far away from the RRHs to save resources. When the value of penalty becomes higher, to avoid higher penalty, the tenant may decide to accept the service requests from users who are far from the RRHs. This may lead to higher resource reservation cost and low profit. If the tenant rejects the service requests from users far away from RRHs to save resources, it will incur a penalty and may also reduce the profit. Our proposed scheme is less sensitive to the value of penalty when it is higher than \$0.004 or lower than \$0.001. This is because when the penalty is low, it plays a negligible role in the profit optimization framework. On the other hand, when the penalty is higher than a certain value, the penalty plays as a dominant role in profit maximization. In this case, most of the service requests from users may be accepted and it is not necessary to further increase the penalty.

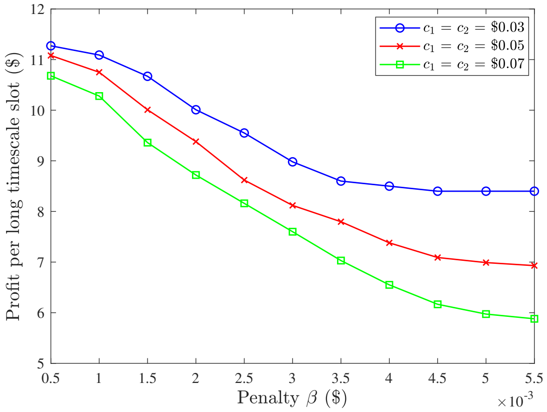


Fig. 14: Profit with different values of resource reservation costs c_1 and c_2, and penalty β, $r^{\text{req}} = 1.5$ Mb/s, $\bar{\chi} = 3$ (number of users per short timescale slot), $\bar{\varepsilon}^2 = 0.05$.

V. Conclusion

In this paper, we proposed a two-timescale framework for resource reservation and intra-slice resource allocation, to maximize the profit of the tenant while guaranteeing the QoS requirements under the CSI uncertainty and user traffic variation. The problem was formulated as a two stage stochastic programming problem. We transformed the stochastic programming problem to a deterministic optimization problem, and applied branch-and-bound and primal-relaxed dual techniques to solve the problem. We evaluated the properties of the designed algorithm to show that the algorithm can solve the problem efficiently. Numerical results indicated that the proposed scheme can well adapt to the variation of user traffic to maximize the profit. By taking into account the CSI uncertainty, the resource reservation and intra-slice resource allocation can guarantee QoS even though real-time channel vector varies. For future work, we will study techniques that can further increase the number of supported users. Examples include the use of millimeter wave frequency band and non-orthogonal multiple access (NOMA) technique.

References

[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] 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.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] 3GPP TS 23.501 V16.3.0, “System architecture for the 5G system; stage 2 (Release 16),” Dec. 2019.

[11] 3GPP TS 28.530 V16.1.0, “Management and orchestration; concepts, use cases and requirements (Release 16),” Dec. 2019.

[12] 3GPP TR 28.801 V15.1.0, “Study on management and orchestration of network slicing for next generation network (Release 15),” Jan. 2018.

[13] 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*, Hong Kong, China, Aug. 2013.

[14] 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.

[15] 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.

[16] 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.

[17] 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.

[18] 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.

[19] 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.

[20] 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.

[21] 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.

[22] 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.

[23] 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.

[24] P. Rost, C. Mannweiler, D. S. Michalopoulos, C. Sartori, V. Sciancalepore, N. Sastry, O. Holland, S. Tayade, B. Han, D. Bega, D. Aziz, and H. Bakker, “Network slicing to enable scalability and flexibility in 5G mobile networks,” *IEEE Communications Magazine*, vol. 55, no. 5, pp. 72–79, May 2017.

[25] J. Zheng, P. Caballero, G. de Veciana, S. J. Baek, and A. Banchs, “Statistical multiplexing and traffic shaping games for network slicing,” *IEEE/ACM Transactions on Networking*, vol. 26, no. 6, pp. 2528–2541, Dec. 2018.

[26] X. Ge, J. Ye, Y. Yang, and Q. Li, “User mobility evaluation for 5G small cell networks based on individual mobility model,” *IEEE Journal on Selected Areas in Communications*, vol. 34, no. 3, pp. 528–541, Mar. 2016.

[27] 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.

[28] 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.

[29] W. Fang, X. Yao, X. Zhao, J. Yin, and N. Xiong, “A stochastic control approach to maximize profit on service provisioning for mobile cloudlet platforms,” *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 48, no. 4, pp. 522–534, Apr. 2018.

[30] M. Dong, X. Liu, Z. Qian, A. Liu, and T. Wang, “QoE-ensured price competition model for emerging mobile networks,” *IEEE Wireless Communications*, vol. 22, no. 4, pp. 50–57, Aug. 2015.

[31] A. Kiani and N. Ansari, “Toward hierarchical mobile edge computing: An auction-based profit maximization approach,” *IEEE Internet of Things Journal*, vol. 4, no. 6, pp. 2082–2091, Dec. 2017.

[32] 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.

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

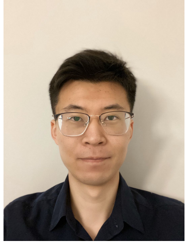
[34] Z. Wang, D. W. K. Ng, V. W. S. Wong, and R. Schober, “Robust beamforming design in C-RAN with sigmoidal utility and capacity-limited backhaul,” *IEEE Transactions on Wireless Communications*, vol. 16, no. 9, pp. 5583–5598, Sept. 2017.

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

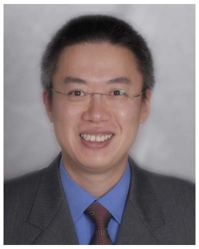
[36] Z. Luo, W. Ma, A. M. So, Y. Ye, and S. Zhang, “Semidefinite relaxation of quadratic optimization problems,” *IEEE Signal Processing Magazine*, vol. 27, no. 3, pp. 20–34, May 2010.

[37] 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.

[38] A. Liu, V. K. N. Lau, and M. Zhao, “Stochastic successive convex optimization for two-timescale hybrid precoding in massive MIMO,” *IEEE Journal of Selected Topics in Signal Processing*, vol. 12, no. 3, pp. 432–444, Jun. 2018.



He Zhang received the B.E. degree from Xi’an Jiaotong University, China, in 2016 and the M.A.Sc. degree from the University of British Columbia, Vancouver, Canada, in 2019. His research interests include machine learning for wireless communications and radio resource allocation. Mr. Zhang is currently working at the New Oriental Education & Technology Group in Beijing, China.



Vincent W.S. Wong (S'94, M'00, SM'07, F'16) received the B.Sc. degree from the University of Manitoba, Winnipeg, MB, Canada, in 1994, the M.A.Sc. degree from the University of Waterloo, Waterloo, ON, Canada, in 1996, and the Ph.D. degree from the University of British Columbia (UBC), Vancouver, BC, Canada, in 2000. From 2000 to 2001, he worked as a systems engineer at PMC-Sierra Inc. (now Microchip Technology Inc.). He joined the Department of Electrical and Computer Engineering at UBC in 2002 and is currently a Professor. His research areas include protocol design, optimization, and resource management of communication networks, with applications to wireless networks, smart grid, mobile edge computing, and Internet of Things. Currently, Dr. Wong is an Executive Editorial Committee Member of *IEEE Transactions on Wireless Communications*, an Area Editor of *IEEE Transactions on Communications*, and an Associate Editor of *IEEE Transactions on Mobile Computing*. He is a Technical Program Co-chair of the *IEEE* 92*nd Vehicular Technology Conference (VTC*2020-*Fall)*. He has served as a Guest Editor of *IEEE Journal on Selected Areas in Communications* and *IEEE Wireless Communications*. He has also served on the editorial boards of *IEEE Transactions on Vehicular Technology* and *Journal of Communications and Networks*. He was a Tutorial Co-Chair of *IEEE Globecom*'18, a Technical Program Co-chair of *IEEE SmartGridComm*'14, as well as a Symposium Co-chair of *IEEE ICC*'18, *IEEE SmartGridComm* ('13, '17) and *IEEE Globecom*'13. He is the Chair of the IEEE Vancouver Joint Communications Chapter and has served as the Chair of the IEEE Communications Society Emerging Technical Sub-Committee on Smart Grid Communications. He is an IEEE Communications Society Distinguished Lecturer (2019 - 2020).