IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received February 16, 2019, accepted March 7, 2019, date of publication March 13, 2019, date of current version April 1, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2904766

Profitable and Energy-Efficient Resource Optimization for Heterogeneous Cloud-Based Radio Access Networks

TAEWOON KIM[1], (Member, IEEE), AND J. MORRIS CHANG[2], (Senior Member, IEEE)
[1]School of Software, Hallym University, Chuncheon 24252, South Korea
[2]Department of Electrical Engineering, University of South Florida, Tampa, FL 33647, USA

Corresponding author: Taewoon Kim (taewoon@hallym.ac.kr)

This research was supported by Hallym University Research Fund, 2018 (HRF-201809-012).

ABSTRACT As network operators invest more and more in infrastructure to keep up with the ever-increasing traffic demand, it has become important for them to operate networks in a profitable manner. The resulting expansion of network infrastructure also increases power consumption, which has a negative impact on both the environment and revenue. In this regard, the cloud-based radio access network (C-RAN), which is a promising next-generation network architecture, has gained much attention as a solution. In addition, by utilizing the macro base stations for coverage, the resulting heterogeneous C-RAN (H-CRAN) can further help optimize the network while increasing the complexity of resource optimization. In this paper, we study an optimal resource allocation for C-RANs to maximize profit while minimizing power consumption considering the inherent network uncertainties. Moreover, by allowing network operators to share networking resources among themselves, we show that the service outage can be minimized without installing additional bandwidth or base stations. The proposed multi-stage stochastic programming model makes a robust optimal decision that effectively responds to uncertainties from users' mobility and service demand while maximizing both profit and energy efficiency. The extensive evaluation and comparison results show that the proposed solution can maximize profit and energy efficiency while minimizing the service outage under network uncertainties.

INDEX TERMS Cloud-RAN, heterogeneous C-RAN, profit maximization, energy efficiency, resource optimization, stochastic programming.

I. INTRODUCTION

THE increase in both the number of smart devices and the volume of network traffic has been served or handled by providing more resource to users, e.g., securing more bandwidth or installing additional base stations (BSs), or by introducing a new air interface standard. However, such an effort causes a huge investment, not to mention replacing most of the network infrastructure if the current hardware does not comply with the new standard. In order to avoid such wasteful overhead, the new radio access network is recommended to meet such requirements as the support of different communication standards without replacing equipments and fulfilling the ever-increasing demand from the never-decreasing number of smart devices in a cost-effective, scalable manner.

The associate editor coordinating the review of this manuscript and approving it for publication was Xiaofan He.

As we move towards the next generation network (5G), we have seen some meaningful progress made to meet the aforementioned requirements. Among those, the centralized or cloud-based radio access network (C-RAN) [1]–[4] which is a completely new cellular network architecture is considered to be one of the key enablers. As shown in Fig. 1, C-RAN consists of a baseband unit (BBU) pool, remote radio heads (RRHs) and fronthaul links that connect them via high-speed links such as optical transmission networks [3]. C-RAN shifts the majority of the functionalities from BSs to the central cloud computing resource pool (i.e., BBU pool), which is in contrast to the conventional configuration where BSs are in charge of MAC/PHY functions. To this end, the BBU pool in the cloud platform performs most of the tasks, e.g., baseband signal processing and transmission scheduling, making RRHs low-complex and low-cost. By virtue of the software-defined radio and network function virtualization,

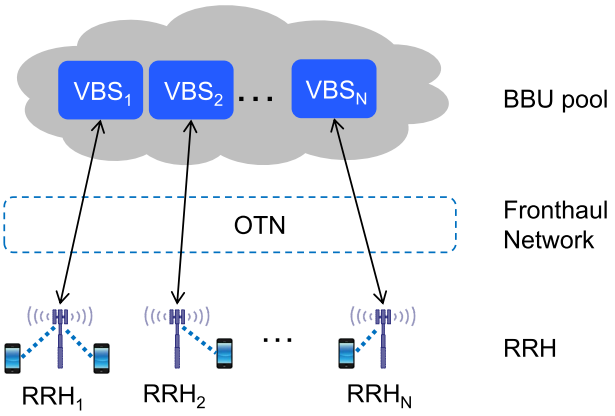

FIGURE 1. C-RAN consists of BBU pool, RRHs and fronthaul links, where optical transport network (OTN) is used to implement the fronthaul network.

C-RANs can become more scalable, reusable and easily configurable. The centralization enables an enhanced cooperation/coordination between RRHs. In addition, the low-cost RRHs allow mobile network operators (MNOs) to densely deploy them (i.e., close to end-users) so that high transmission rates can be achieved with low transmission power. In the long run, C-RAN is expected to reduce the total cost of ownership, i.e., the sum of capital expenditure (CAPEX) and operational expenditure (OPEX), which makes it more attractive [3]. Moreover, the use of short-range, low-power RRHs helps enhance energy efficiency.

In particular, the heterogeneous C-RAN (H-CRAN) [5] which is an advancement to or variation on C-RAN is more attractive for the following reasons. By retaining the use of the conventional long-range macro base stations (MBSs) on top of C-RAN, H-CRANs can lead to a gradual transition from 4G to 5G. MBSs and RRHs, respectively, can be dedicated to control message exchanges and high-speed data transfer, separating the control plane from data in a straightforward manner. In addition, due to the large coverage provided by MBSs, there is no need to densely deploy RRHs as shown in Fig. 2. In fact, densely-deployed RRHs may cause control message overflow (e.g., due to frequent handovers [3]), increase the complexity of the interference management, and leave many of them under-utilized [6]. Such a heterogeneous structure in H-CRAN, however, gives rise to a new challenge for the increased complexity at the same time.

A. CHALLENGES AND MOTIVATION

The biggest challenge in both C-RAN and H-CRAN is that the set of resources to consider has become larger and not all can be treated the same in terms of the timescale of resource scheduling (or how frequently each resource can be and should be scheduled). Optimizing H-CRAN is, in general, followed by not only the conventional tasks (e.g., user-BS association and channel assignment), but also the new ones introduced by the cloud-based architecture (e.g., computing resource allocation) and new constraints (e.g., fronthaul capacity limitation). Since adjusting one affects the

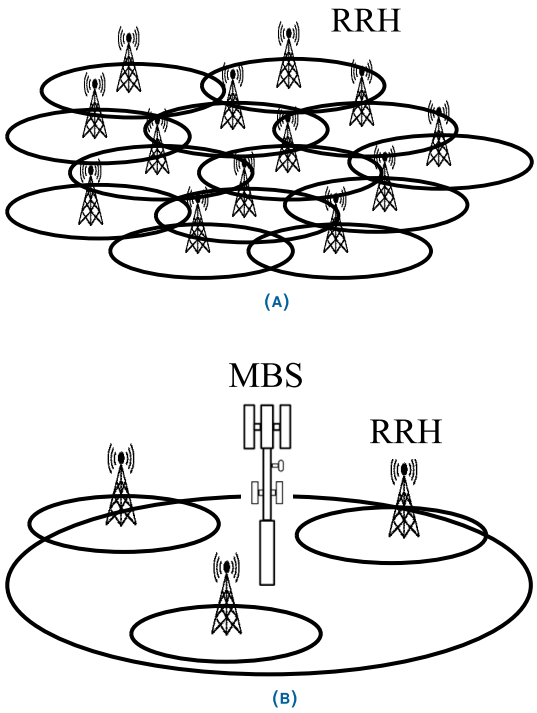

FIGURE 2. Comparison between C-RAN and H-CRAN. (a) In C-RAN, RRHs are densely deployed to provide coverage, causing severe interference. (b) In H-CRAN, MBSs provide coverage, while heavy traffic is offloaded to RRHs.

rest, and as a result, limits the overall network performance, it is nearly impossible to truly optimize the new RAN without taking into account the entire resource *chain* present in H-CRAN.

In addition, despite of the importance of different timescale issue in resource optimization, it has not been received much attention. For conventional cellular networking systems, resource scheduling (e.g., channel assignment, transmission power control and user-BS association) is carried out by the MBSs to which UEs (user equipments)[1] are directly attached at the moment. Since BSs make decisions locally, scheduling such resources can be done without any significant delay. In C-RAN/H-CRAN, on the other hand, BS functionalities are shifted to the BBU, forming Virtual Base Stations (VBSs). VBSs are built on the cloud platform by means of Virtual Machines (VMs), resulting in both the increased system utilization and the decreased CAPEX [7]. The capacity of a VBS has a significant impact on the performance of the associated RRH, and thus it needs to be carefully determined. Given that resizing and reconfiguring VBSs triggers re-partitioning of BBU which is a time-consuming task,[2] frequent VBS adjustment or scheduling of BBU may bring a non-negligible service delay. In this regard, when scheduling

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

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

C-RAN/H-CRAN resources, such time-consuming tasks need to be identified and separately handled from the ones that can be scheduled at short intervals; otherwise, users may suffer from frequent service outage.

Another challenge that motivated this study is related to the heterogeneous architecture of H-CRAN as well as the inherent network uncertainties. As reported in [8], the traffic load greatly changes over time and space due to the fluctuating users' demand [9] and their mobility. When there are short-range BSs (i.e., RRHs) deployed, the connectivity (or accessibility) between users and such BSs is largely affected by their geographic locations at the moment. Since such BSs can provide high-bandwidth with low-power, it is significantly important to aware of the users' mobility and the varying service demand so that RRHs can be actively utilized by means of traffic offloading. Thus, to satisfy users demand in an energy-efficient manner, it is important to take the uncertainties in both mobility and service demand into account at the same time for resource scheduling, which has not been carefully studied yet.

Nevertheless, it is almost impossible to guarantee a zero service outage at all times. For example, there can be a sport or social event for which a lot of people gather together in a certain area. The service demand from the region will explode, and as a result, the users in the region will experience service quality degradation. It is an important task for network operators to effectively handle such sporadic demand bursts in a cost-efficient manner. This is because the conventional approaches to preparing such rare, worst-case scenarios, e.g., installing more BS, may result in a high expenditure and energy consumption, imposing a negative impact on customers (e.g., increasing service fee) and environment, respectively. Thus, from the network operator's point of view, it is important to optimize the network resources to maximize both the revenue and energy efficiency without negatively affecting service quality. As pointed out in [3], this is an increasingly important mission especially as the network traffic increases.

In contrast to the conventional resource optimization approaches focusing on either throughput maximization or energy minimization from the users perspective, only a few studies have paid attention to profit maximization from the operator's perspective [10]–[13]. It has recently been noticed that compared to the increasing investment MNOs make to upgrade or scale out their network infrastructure, the average revenue per user has been small [3], [13]. Also, the power consumption from the expanding infrastructure accounts for both increasing OPEX and a large share of carbon footprint in the information and communications technology domain. In this regard, the energy-efficiency is perceived as an important performance metric in 5G networking systems [14]. Thus, there is an urgent need for studying a profitable and energy-efficient operation of the networking system in preparation for the upcoming 5G.

B. PROPOSED SOLUTION AND CONTRIBUTION

In this paper, we propose an optimal resource scheduling method for H-CRAN under network uncertainties. In order for an MNO to maximize the profit, the proposed method optimizes the comprehensive set of H-CRAN resources, including the partitioning of BBU pool for RRHs, the channel partitioning between different types of BSs, the association between BSs and users, and the channel assignment for each user. The fronthaul constraint and the power consumption are taken into account to make the proposed model to be more practical and energy-efficient, respectively. In order to comply with the overall goal in this study, we propose a new energy efficiency metric in which the profit-energy relation is embedded. The new energy-efficiency metric is defined as the amount of revenue per energy consumption. The metric implies how much revenue a network operator can make by consuming a unit energy. This definition is different from the conventional ones, such as the achieved data rate per energy consumption [14], [15].

To compute the optimal resource scheduling strategy, we propose to apply a stochastic programming (SP) approach. By doing so, the proposed model can effectively react to the uncertainties stemming from the users' mobility and their varying service demand. In addition, by leveraging the staged structure of SP, the proposed model finds one-fits-all, robust long-term solutions for the resources that should be scheduled at long intervals. For the rest resources, on the other hand, it finds optimal short-term solutions in an online manner. In the other studies, the average values for uncertain parameters are assumed. However, such approaches works only for numeric analysis, and not applicable for real scenarios or even simulations.

Also, we propose an effective way to achieve a profitable cooperation between different MNOs by allowing users to offload their traffic to other MNOs [10], [11], [16], [17] when there is not enough resource left in their primary MNO's network. By an extensive evaluation, we compare the proposed model to others that do not or partially consider the uncertainties, and show that the proposed model can achieve a higher profit than the rest while minimizing both the power consumption and service outage ratio. Also, we show the promising effect allowing a cooperation among different MNOs on the profit.

The major contributions we make in this study are summarized below.

- We formulate an extensive resource optimization problem for H-CRAN considering the comprehensive resources related to providing network service to users. The proposed model is aware of the different timescales in resource scheduling, which is critical to scheduling cloud-based RAN resources.
- We propose to use a SP-based approach to effectively handle the network uncertainties, which is crucial to provide a practical solution especially for H-CRANs with heterogeneous BSs.

TABLE 1. Summary of related work.

Topic	Paper(s)
profit, cost	[13], [26]
power, energy	[9], [13], [14], [18], [20], [22], [25], [27], [28], [29], [30], [31]
throughput, utilization, fairness	[7], [19], [23], [24]
service quality	[18], [19]
social welfare	[21]

- We propose a cost-and-profit model for the cloud-based RANs, with which we study how to maximize the overall profit of the network operator as well as the energy efficiency. To do so, we also propose a new energy-efficiency metric which implies the revenue-energy relation.
- We propose a profitable way to utilize the networking resources of other network operators if the network becomes saturated. By doing so, a network operator can minimize the service outage without expanding the network capacity.
- We design and carry out a set of extensive evaluations to verify the effectiveness of the proposed solution approach. We also implement the baseline method as well as its variations to compare the performance of the proposed method.

C. ORGANIZATION

The remainder of this paper is organized as follows. Section II summarizes related literatures. Section III introduces the overview of the problem, network model and essential elements in formulating an SP. In Section IV, we formulate the optimal resource allocation problem by means of multi-stage SP. Section V presents the evaluation results, and finally, Section VI concludes this paper.

II. RELATED WORK

The concept of the radio access network with cloud computing was first proposed by IBM [1] and then, took the shape of C-RAN by China Mobile in 2011 [2]. Later, C-RAN with a heterogeneous network architecture, called H-CRAN [5], was proposed to further enhance the spectral and energy efficiency. This section introduces the related works to C-RAN or H-RAN resource optimization. The main focus of each study to be discussed in this section is summarized in Table 1.

Wang *et al.* [13] formulated a profit maximization problem for C-RAN with mobile cloud computing. While maximizing the serviced traffic, their proposed method minimizes the energy consumption of both fronthaul and server. However, optimal server partition and networking resource scheduling (e.g., channel assignment between UEs and RRHs) are not considered. Luo *et al.* [9] studied the C-RAN where access points (APs) are densely deployed. In order to reduce the interference and thus to increase the energy efficiency, they proposed an optimal scheduling scheme for user association, AP activation and beamforming in downlink and uplink transmission. Pompili *et al.* [7] proposed a demand-aware resource provisioning scheme that dynamically resizes the VBSs so as to enhance resource utilization and system performance. The first component of the proposed scheme, called, proactive component, pre-allocates VBSs given the historical traffic profile. During operation, when a significant mismatch happens between the anticipated and actual traffic, the reactive component adjusts the VBSs.

Lyazidi *et al.* [18] proposed a two-stage resource allocation scheme for C-RAN. Considering the traffic fluctuations, the proposed scheme dynamically allocates the frequency resource to UEs, and makes associations between RRHs and BBUs with respect to quality of service (QoS) requirements and energy minimization. Specifically, mixed integer programming and Knapsack problem are used to formulate the problem. Feng *et al.* [19] proposed a resource allocation algorithm to increase both QoS and fairness for the public safety network on C-RAN. In particular, the authors proposed a resource block allocation algorithm between users and RRHs. To make the algorithm efficient, authors relaxed the integer variables and then recovered integer solution by using Feasible Pump method. Tang *et al.*1 [20] proposed a cross-layer approach to minimize the power consumption of C-RAN system, i.e., BBU, fiber links and RRHs. To this end, the authors formulated optimization problems for VM allocation, RRH selection and beamforming strategies. For the solution method, extended sum-utility maximization and Shaping-and-Pruning are used.

Gu *et al.* [21] proposed a model where network operators can lease resource from tower company by using an auction mechanism. In their problem setting, a tower company sells spectrum and bandwidth resource to network operators, which is to be solved by the proposed near-optimal resource provisioning algorithm with maximizing the social welfare. Zhao and Wang [22] investigated the power consumption in C-RANs focusing on both RRHs and an optical network (i.e., a fronthaul network). Given the traffic density in the service region, the proposed method chooses a subset of RRHs to activate under constrains including power/bandwidth budget of RRHs and QoS of UEs.

Cai *et al.* [23], [24] studied the topology configuration and rate allocation in C-RANs in mobile cloud computing systems. Given the delayed channel state information, they formulated an optimization problem to maximize the transmission control protocol (TCP) end-to-end throughput. Mashayekhy *et al.* [25] proposed an auction-based mechanism which provisions the virtual machines to physical machines, and then assigns them to users aiming at energy efficient resource management. Chaisiri *et al.* [26] formulated a two-stage stochastic programming to provision the central computing resources under demand and price uncertainties aiming at minimizing the provisioning cost. The proposed approaches therein can be used to optimize the cloud computing resource in a general setting, while not applicable to C-RAN or H-CRAN.

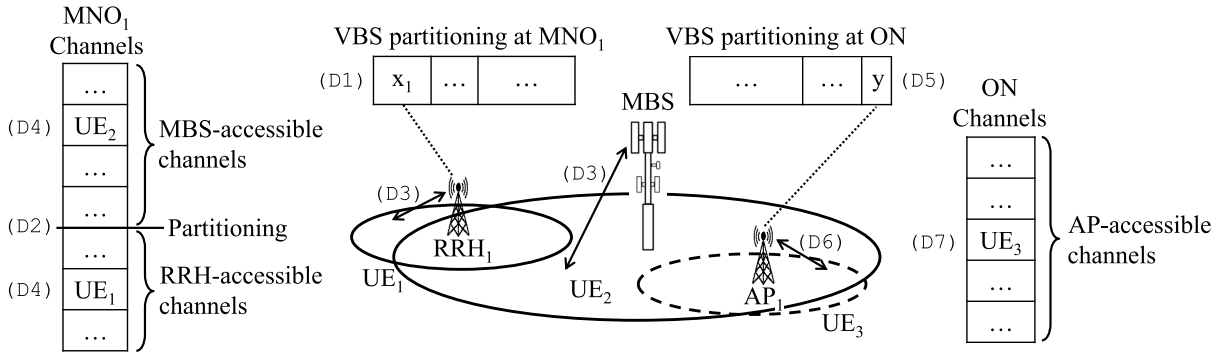

FIGURE 3. Illustration of the proposed resource scheduling method, highlighting the set of decisions to be made from (D1) to (D7). All UEs shown in the figure are MNO_1's subscribers.

The power minimization problem in [27] utilizes the cooperative transmission among RRHs and dynamic VM assignment for UEs under limited fronthaul capacity. Also, queuing-based resource optimization problems were formulated in [28] and [29], respectively, to study the throughput-delay trade-off and to maximize the energy efficiency. Liu *et al.* [30] modeled and formulated the network-wide energy efficiency problem in H-CRAN. The proposed model therein captures the power consumption from base stations, fronthaul and the BBU pool. The initial mixed integer non-linear programming model is transformed into a computationally efficient algorithm, called HERM. Their work, however, lacks considering the network uncertainty. In addition, the optimal VBS partitioning is missing, which may not lead to the truly optimal resource use of H-CRANs.

Pan *et al.* [14] formulated a joint precoding and RRH selection for C-RANs. The proposed 2-stage low-complex method, in turn, maximizes the number of admitted users and then solves a power minimization problem. The same author proposed a user selection and power minimization for C-RAN with incomplete channel state information in [31]. In addition to the dense C-RANs they considered in those studies, other differences compared to the present paper is that their work lack in considering both optimizing the BBU pool and the network uncertainties.

Compared to the previous works, the proposed method in this paper is unique and novel as follows. First of all, we consider the entire H-CRAN resources instead of focusing on a partial subset. Although it significantly increases complexity in problem formulation, it is significantly important because scheduling one resource limits the others and the overall performance significantly. Thus, it is critical to identify and optimize the complete set of networking resources for the 5G H-CRAN network. Secondly, none of the previous works have jointly considered the inherent network uncertainties in both users' mobility and their fluctuating service demand. However, such uncertainties have a significant effect on the resource optimization especially when there are short-range base stations are deployed. To propose a practical solution, we consider the both uncertainties in this work. Thirdly, for the first time, we pay attention to the different resource scheduling intervals to be considered when scheduling both H-CRAN and C-RAN resources. This is a new challenge we identified in the cloud-based network resource optimization problem. In addition, we propose a profit-maximizing resource optimization, which is the goal of any network operators. By allowing traffic offloading to different network operators, the proposed solution can further enhance the network capacity without acquiring addition network resources.

III. PROBLEM DESCRIPTION

A. PROBLEM OVERVIEW

In this paper, we study optimal resource scheduling for H-CRAN under uncertainty. The objective is to maximize profit while minimizing power consumption. We assume that there are multiple MNOs operating in the given region, and we focus on MNO_1 along with its resources and subscribers. MNO_1's network resources as well as its operating frequencies are independent of that of the other MNOs. In addition, we assume a special type of MNO, referred to as Open Network or ON for short, whose business model is different from ordinary MNOs. ON has a pool of network resources that can be leased and used by other MNOs by means of traffic offloading. In order to avoid confusion in naming different type of BSs, we use the following notations:

- **MBS** for high-power, long-range macro base station operated by MNO_1,
- **RRH**s for remote radio heads operated by MNO_1, and
- **AP**s for access point[3] operated by ON, and they are also accessible to MNO_1's subscribers if both MNO_1 and ON agreed to do so beforehand.

In what follows, *base station (BS)* refers to any or all of the above three.

In the proposed method, MNO_1 makes the following seven decisions from (D1) to (D7) as illustrated in Fig. 3. (D1) MNO_1 partitions its BBU pool (i.e., the central computing resource), and then, forms VBSs and assigns them to RRHs.

[3]Please note that an AP can be RRH, MBS, low-power BS, 802.11-type AP, or something else depending on the ON's network model.

As aforementioned, one VBS is assigned to one RRH. Also, (D2) the available, orthogonal channels are divided into two disjoint sets, one for MBS and the other for RRHs; as a result, there is no cross-tier interference. During operation, UEs receive service from either MBS or RRH after both (D3) UE-BS association and (D4) the channel assignment for each UE are determined. In addition, MNO_1 is allowed to collaborate with ON in a way that MNO_1's subscribers can offload their traffic to ON via its APs. To do so, (D5) MNO_1 has to reserve a certain amount of computing resource at ON in advance. For MNO_1, when the aggregate service demand exceeds the network capacity, (D6) it allows UEs to offload their traffic to ON's APs, and at the same time (D7) which channel to use for such user is determined. The aggregate profit to MNO_1 is revenue minus the sum of both cost and penalty. The revenue increases in the subscribers' data usage, i.e., number of bits transferred and processed. On the other hand, MNO_1 has to pay the cost when it reserves computing resource at ON (i.e., resource reservation fee) and when its subscribers offload their traffic to ON's APs (i.e., offloading cost). Also, penalty is charged to MNO_1 for any unmet demand.

In order to effectively respond to the network uncertainties, i.e., users' mobility and their service demand, we formulate an resource optimization problem in SP [32], [33]. SP is a mathematical programming tool in which some of the parameters are uncertain and described by their probability distributions. In contrast to previous C-RAN resource optimization approaches that eliminate uncertainties by assuming constant values (i.e., expectations) for uncertain parameters, SP yields a more robust and practical solution since it takes possible realizations of such parameters into account. As a result, the problem of resource over-/under-provisioning is minimized. Also, both the separation of stages and recourse actions[4] [32] in SP result in a structured problem formulation and robust solutions against uncertainties, respectively. In this paper, *realizations* of an uncertainty parameter refer to the possible values that the parameter can take on, and *scenario* refers to the set of such realizations that can occur at the same time. One scenario includes multiple realizations, one for each uncertainty parameter, and the set of such scenarios is called scenario tree.

B. NETWORK MODEL AND ASSUMPTIONS

On the network is an MBS overlaid by $|\mathcal{M}_1|$ RRHs. The network is operated by MNO_1 which has $|\mathcal{U}|$ number of active UEs or subscribers. Low-power RRHs are sparsely deployed[5] in a planned manner such that there is no intra-tier interference. Due to the low transmit power of RRHs [2] as well as their sparsity in distribution (i.e., RRHs are distributed far apart from each other), an RRH does not interfere with the rest of the same kind. Therefore, in this work, the entire spectrum that is accessible to RRHs can be reused with factor of 1. Each RRH is directly connected to its serving VBS in the BBU pool via a wired, capacity-limited fronthaul link [34]. The MBS provides coverage, meaning that each UE can always access it; yet, it is not always the case between UEs and RRHs. On the same region, ON has its own APs installed. Both operators have a limited number of non-overlapping orthogonal channels each, and thus, there is no inter-network interference. Each UE has two radio interfaces [14], one for MNO_1 and the other for ON. A UE can only be associated with a single MBS or RRH at the same time over the radio for MNO_1, and which is the same for AP over the radio for ON.[6] Service demand (i.e., downlink traffic rate) of each UE fluctuate while following the uniform random distribution with a known mean value.

We assume that MNO_1 has an access to historical data (or has a knowledge of known patterns [7]) with respect to UEs mobility. In particular, what we are interested in is the *accessibility* information between UEs and BSs. During operation, a UE periodically listens to the pilot signals from nearby BSs and transmits the list of accessible BSs (i.e., MBS, RRH or AP) in the uplink. After accumulating such information, an MNO computes the probability that a UE is accessible to a particular BS during a certain period of time, e.g., a day. We also assume an equal per-channel transmission power at a BS, which provides close-to-optimal performance [8], [37]. The BBU pool is assumed to be continuous, and fractional partitioning is allowed. A VBS is assigned to a single RRH to process baseband signals for the RRH [27], [30], [35].[7]

Notations are summarized in Table 2, while the others are defined as needed. For indexing purpose, u, i_1, i_2, m_1, m_2 and s are used to indicate a UE, RRH, AP, MNO_1's channel, ON's channel and scenario, respectively. Vectors are in bold and lowercase, e.g., $\mathbf{x}$, $\boldsymbol{\pi}$ and $\mathbf{1}$. Matrices are in uppercase, bold letters, e.g., $\mathbf{A}_s$ and $\mathbf{B}_s$. Uppercase letters in calligraphic font indicate sets e.g., $\mathcal{I}_1$ and $\mathcal{M}_2$.

C. DECISIONS AND TIMELINE

Time is slotted as shown in Fig. 4. The proposed resource optimization problem is composed of three problems, referred to as Stage-1, Stage-2 and Stage-3. The Stage-1 is to optimize the long-term network resources that should not be frequently scheduled. On the other hand, both Stage-2 and Stage-3 are to schedule the rest network

[4]Please note that *recourse* action in the context of SP refers to an action to be taken in response to each random outcome, and is not a typo.

[5]In H-CRAN, the coverage is provided by MBS while high spectral efficiency can be achieved by RRHs. In fact, a dense deployment of low-power BSs may not be efficient since many of them are often under-utilized [6], and it also increases both interference and energy consumption [9].

[6]In a (ultra) dense network, where coordinated multi-point (CoMP) transmission/reception is exercised [36], a UE can associate with multiple RRHs [14]. However, in this work, we assume an MBS-assisted, heterogeneous C-RANs, where RRHs are sparsely deployed. Therefore, a UE can associate with up to a single RRH. In addition, due to the low transmission power of RRHs, we assume there is no intra-tier interference among RRHs.

[7]Please note that it is technically sound to share a VBS among multiple RRHs as well. Such a configuration is inevitable especially when the number of VBSs that can be instantiated is limited and the number of RRHs is greater than that of VBSs to create. In this work, however, we do not assume such a limitation. Also, the proposed scheme can easily apply to such cases by merging some VBSs into one.

TABLE 2. Summary of notations.

	Sets
$\mathcal{I}_1$	Index set of RRHs, $\{1, 2, \cdots, i_1, \cdots\}$
$\mathcal{I}_2$	Index set of APs, $\{1, 2, \cdots, i_2, \cdots\}$
$\mathcal{M}_1$	Index set of channels MNO_1 can access to, $\{1, 2, \cdots, m_1, \cdots\}$
$\mathcal{M}_2$	Index set of channels ON can access to, $\{1, 2, \cdots, m_2, \cdots\}$
$\mathcal{S}$	Index set of scenarios, $\{1, 2, \cdots, s, \cdots\}$, $\lvert\mathcal{S}\rvert = \lvert\mathcal{T}\rvert$
$\mathcal{T}$	Complete scenario tree (= complete set of scenarios)
$\mathcal{U}$	Index set of UEs, $\{1, 2, \cdots, u, \cdots\}$
	Uncertainty parameters
$\mathbf{r}$	Users' service demand
$\mathbf{V}$	Accessibility indicator between UEs and RRHs
$\mathbf{W}$	Accessibility indicator between UEs and APs
ξ	Description of accessibility and demand, $(\mathbf{V}, \mathbf{W}, \mathbf{r})$
	Deterministic parameters
α	Conversion parameter from serviced bits to revenue
β	Offloading cost
c_{i_1}	Fronthaul capacity between RRH i_1 and its serving VBS
d_{i_2}	Association capacity of AP i_2
δ	OPEX of BBU pool or the cloud computing resource
ϵ	Computing resource reservation cost
$f^u_{i_1,m_1}$	Bandwidth of channel m_1 between RRH i_1 and UE u
$g^u_{i_2,m_2}$	Bandwidth of channel m_2 between AP i_2 and UE u
γ	Penalty for unmet demand
K_1	MNO_1's conversion parameter for processing capacity
K_2	ON's conversion parameter for processing capacity
o_{i_1}	Association capacity of RRH i_1
τ^M	Per-channel transmission power of MBS
τ^R	Per-channel transmission power of RRH
θ	Electricity fee
z_{i_2}	Aggregate bandwidth capacity of AP i_2
	Decision variables
x_{i_1}	Fraction of BBU pool allocated to RRH i_1, where $x_{i_1} \in \mathbf{x}$
y	Fraction of computing resource reserved at ON's BBU pool
$\boldsymbol{\pi}$ or $\overline{\boldsymbol{\pi}}$	MNO_1's channel allocation vector for MBSs or RRHs
$\mathbf{A}_s$	UE, BS (MBS or RRH) and channel mapping in scenario s
$\mathbf{B}_s$	UE, AP and channel mapping in scenario s
$\widetilde{r^u_s}$	Unmet demand of UE u in scenario s, where $\widetilde{r^u_s} \in \widetilde{\mathbf{r}}_s$

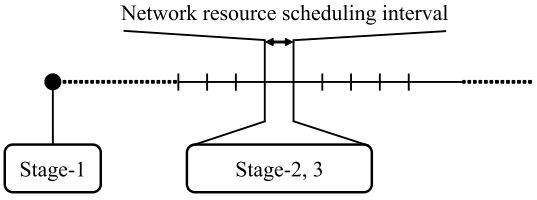

FIGURE 4. In this work, we assume a slotted time frame. At the beginning, the `Stage-1` carries out a long-term resource allocation. Afterwards, for each short-length time slot, both `Stage-2` and `Stage-3` decisions are made to provide network service to users at short intervals.

resources at short intervals to best adapt to network dynamics. The set of resources that needs to be scheduled at short- or long-term intervals will be explained shortly.

At the beginning, the `Stage-1` problem is executed, while both `Stage-2` and `Stage-3` problems run during each time slot. The first stage, `Stage-1`, is for network planning, `Stage-2` is for service provisioning, and the last stage, `Stage-3`, is for a recourse action. At `Stage-1`, without knowing both the UEs' locations and their service demand, MNO_1 makes the following long-term decisions that will remain the same for a while. The BBU pool is partitioned to form VBSs for RRHs, i.e., `(D1)`. Since VBSs are formed by means of VMs, reconfiguring or resizing VMs can cause a significant delay which might be intolerable for realtime cellular networking service. Channel partitioning, i.e., `(D2)`, also belongs to `Stage-1` since it may cause a significant delay for the heavy computation load which will be discussed in Section IV. Also, reserving ON's computation resource by lease, i.e., `(D5)`, should be done in advance at `Stage-1`. This is because ON's availability changes over time, and on-demand request without any in-advance reservation may not be accepted if ON is saturated at the moment. Aforementioned decisions can be made at long intervals.

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

D. UNCERTAINTIES AND SCENARIO REPRESENTATION

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

In SP, uncertain parameters are taken into account by means of *scenarios*. In a particular scenario s, the network snapshot is known to be $\xi_s = (\mathbf{V}_s, \mathbf{W}_s, \mathbf{r}_s)$, where $\mathbf{V}_s$, $\mathbf{W}_s$, $\mathbf{r}_s$ are the realizations drawn from their corresponding distributions. A scenario is a realized snapshot of the network, and it can occur at any time instance. The set of such scenarios constitutes a *scenario tree*. A complete scenario tree, $\mathcal{T}$, is a Cartesian product of all possible realizations of all uncertain parameters, describing all possible snapshots that can occur at any time instance. Since the continuous random variable $\mathbf{r}$ has an infinite number of realizations, resulting in an infinite number of scenarios, we use a discretization method proposed by [48] for $\mathbf{r}$ to reduce the problem size, and thus, to be able to

find the optimal solution in a finite time; it will be discussed in both Section IV-C and Section V-B.

IV. PROBLEM FORMULATION

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

In a nutshell, the joint channel partitioning and channel assignment problem can be (excessively) simplified as follows. Let us assume there are one MBS and one RRH on the network. Let $x_n \in \mathbf{x}$ be a binary decision variable indicating whether a channel n is available to MBS or RRH with x_n being 1 or 0, respectively. Let $y_n^u \in \mathbf{y}$ be a binary decision variable indicating whether a UE u is associated with channel n. Let $m_n^u \in \mathbf{m}^u$ and $r_n^u \in \mathbf{r}^u$ be the data rate UE u can get from MBS and RRH, respectively, if the UE is associated with channel n. If UE u has no access to the RRH, we have $r_n^u = 0$. Then, the actual data rate that UE u can get over all channels is expressed as $\sum_{\forall n} x_n y_n^u m_n^u + \sum_{\forall n}(1 - x_n) y_n^u r_n^u$, which should be greater than or equal to the demand of UE u. Due to the coupling of the decision variables x_n and y_n^u, the problem becomes nonconvex. To make the problem convex, the proposed decomposition and iterative search method takes the following approach. The proposed method first makes a decision on $\mathbf{x}$, the channel partitioning, alone. Then, the proposed problem solves the optimization problem with the rest decision variables, while taking $\mathbf{x}$ as a constant. The proposed problem repeats over different $\mathbf{x}$'s or channel partitions, and in the end, returns the best among all trials.

Algorithm 1 Iterative Resource Allocation Algorithm

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

A. DECOMPOSITION AND ITERATIVE SEARCH

The proposed iterative resource allocation algorithm is given in Algorithm 1. Both ρ^* and ρ^t indicate the optimal objective value and the objective value at t-th iteration, respectively. The composite variable Ω is a set of the decision variables we consider in the proposed problem formulation except the channel partitioning decision. Both Ω^* and Ω^t, respectively, denote the optimal decision and the decision found at t-th iteration. Lines 1–3 initialize the optimal objective value, optimal solution set and optimal channel partitioning in sequence. Line 4 initializes the channel partitioning vector $\boldsymbol{\pi}$ to a zero vector. On each iteration of the **for** loop (lines 5–11), the number of channels assigned or made available to MBS increases from 1 to $|\mathcal{M}_1| - 1$ by 1 (line 6), where $\mathcal{M}_1$ is the index set of channels accessible to MNO_1.

The $\boldsymbol{\pi} \in \{0, 1\}^{|\mathcal{M}_1|}$ is a channel allocation vector for MBS, where having 1 or 0 at t-th element $\boldsymbol{\pi}[t]$ indicates that the corresponding channel t is open to use to MBS or not, respectively. We assume consecutive channel allocation, meaning that the set of channels accessible to MBS is $\{1, 2, \cdots, m_1\}$ whereas that to RRH is $\{m_1 + 1, m_2 + 2, \cdots, |\mathcal{M}_1|\}$. Also, there should be at least one channel available to MBS and RRHs each, for which the **for** loop at line 5 begins and ends with 1 and $|\mathcal{M}_1| - 1$, respectively. The accessible channels to MBS and RRH do not overlap with each other, and thus there is no cross-tier interference. RRHs operate on the channels that are marked by zeros in $\boldsymbol{\pi}$. We denote such channels by $\overline{\boldsymbol{\pi}} = \mathbf{1}_{|\mathcal{M}_1|} - \boldsymbol{\pi}$, where $\mathbf{1}_{|\mathcal{M}_1|}$ is a vector of all 1's with $|\mathcal{M}_1|$ entries therein. As a result, RRH-accessible channels are marked by 1's in $\overline{\boldsymbol{\pi}}$. Given the power budget P_{max}^M for an MBS and P_{max}^R for an RRH, the per-channel transmit power of MBS and RRH becomes $P_{max}^M/|\boldsymbol{\pi}|$ and $P_{max}^R/|\overline{\boldsymbol{\pi}}|$, respectively.

In line 7, Algorithm 1 solves deterministic equivalent problem (DEP), which will be introduced in Section IV-C. Then, we further reduce the problem complexity in Section V-B. In a nutshell, we first form an energy-aware profit maximization problem in SP, which is intractable. Thus, we transform it into a DEP instance that can be solved with a computer solver, e.g., Matlab [38], in a finite time. Lines 8–9 compare the best objective value so far to the current one, and take the one yielding a larger profit. Once the algorithm terminates, it returns both Ω^* and $\boldsymbol{\pi}^*$ (line 12), with which MNO_1 configures its network.

Please note that Algorithm 1 may not produce a global optimal solution if optimum can only be found with a non-consecutive channel partitioning. However, the assumption on consecutive channel allocation helps to reduce the complexity of the entire algorithm significantly. In addition, if the globally optimal solution exists with consecutive channel partitioning, the proposed algorithm can produce the same, global optimal solution.

In what follows, we first formulate three problems P. 1–P. 3, one for each stage, in Section IV-B. Then, in Section IV-C we transform the entire problem into DEP, P. 4, by replacing the expectation term in (1a) with the sum of profit multiplied by the probability of having the corresponding event. Then, in Section V-B the complexity of P. 4 is further reduced by diminishing the problem size (i.e., the number of scenarios to

include in the scenario tree) so that we can compute the optimal solution with a computer solver within a finite amount of time.

B. MULTI-STAGE STOCHASTIC PROGRAMMING

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

$$\max_{\mathbf{x},y} \; -\delta \sum_{i_1\in\mathcal{I}_1} x_{i_1} - \epsilon\cdot y + \mathcal{Q}(\mathbf{x},y) \tag{1a}$$

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

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

$$0 \le y \le 1, \tag{1d}$$

$$\forall i_1: \; x_{i_1}\cdot K_1 \le c_{i_1}, \tag{1e}$$

The objective (1a) is to maximize the total profit, i.e., revenue minus cost. The first term is the OPEX on operating the BBU pool, where $x_{i_1} \in \mathbf{x}$ is the fraction of BBU allocated to RRH $i_1 \in \mathcal{I}_1$, and δ is a conversion parameter from x_{i_1} to expense (\$). The second term is the rental fee, where y is the fraction of computing resource that MNO_1 reserves at ON in advance, and ϵ is a conversion parameter from y to the rental fee (\$). The third term is the expected profit from `Stage-2` given $\mathbf{x}$ and y, i.e., $\mathcal{Q}(\mathbf{x},y) = \mathbb{E}_\xi[Q_s(\mathbf{x},y;\xi_s)]$, where $\xi = (\mathbf{V},\mathbf{W},\mathbf{r})$ is the uncertain parameter describing a network snapshot (i.e., accessibility and demand), $\xi_s = (\mathbf{V}_s,\mathbf{W}_s,\mathbf{r}_s)$ is a realized scenario indexed by s, and $Q_s(\mathbf{x},y;\xi_s)$ is the `Stage-2` objective value given the `Stage-1` decisions (i.e., $\mathbf{x}$ and y) and ξ_s.

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

`Stage-2` is for service provisioning, whose problem formulation is shown below (called P. 2). Given the decisions made in `Stage-1` as well as the scenario that has become known at the beginning of each time slot, MNO_1 schedules its own networking resources to provide service to its subscribers. Please note that out of three realizations, i.e., $\mathbf{V}_s$, $\mathbf{W}_s$ and $\mathbf{r}_s$, that become known, MNO_1 makes use of only both $\mathbf{V}_s$ and $\mathbf{r}_s$ in `Stage-2`. As a reminder, MNO_1 utilizes its own networking resource in `Stage-2`. The $\mathbf{W}_s$ will be used in `Stage-3` if MNO_1 has to offload some traffic to ON.

$$\begin{aligned} Q_s(\mathbf{x},y;\xi_s) := \max_{\mathbf{A}_s} \; & \alpha\Big(\sum_{u\in\mathcal{U}}\sum_{m_1\in\mathcal{M}_1} a^u_{0,m_1,s}\cdot f^u_{0,m_1}\Big) \\ & -\theta\cdot\tau^M \sum_{u\in\mathcal{U}}\sum_{m_1\in\mathcal{M}_1} a^u_{0,m_1,s} \\ & +\alpha\Big(\sum_{i_1\in\mathcal{I}_1}\sum_{u\in\mathcal{U}}\sum_{m_1\in\mathcal{M}_1} a^u_{i_1,m_1,s}\cdot f^u_{i_1,m_1}\Big) \\ & -\theta\cdot\tau^R \sum_{i_1\in\mathcal{I}_1}\sum_{u\in\mathcal{U}}\sum_{m_1\in\mathcal{M}_1} a^u_{i_1,m_1,s} \\ & +H_s(y,\mathbf{A}_s;\xi_s) \end{aligned} \tag{2a}$$

$$\text{subject to}$$

$$\forall u: \; \sum_{m_1\in\mathcal{M}_1} a^u_{0,m_1,s} + \sum_{i_1\in\mathcal{I}_1}\sum_{m_1\in\mathcal{M}_1} a^u_{i_1,m_1,s} \le 1, \tag{2b}$$

$$\forall u,m_1: \; 0 \le a^u_{0,m_1,s} \le \boldsymbol{\pi}_{m_1}, \tag{2c}$$

$$\forall u,m_1,i_1: \; 0 \le a^u_{i_1,m_1,s} \le v^u_{i_1,s}\cdot\overline{\boldsymbol{\pi}}_{m_1}, \tag{2d}$$

$$\forall i_1: \; \sum_{u\in\mathcal{U}}\sum_{m_1\in\mathcal{M}_1} a^u_{i_1,m_1,s} \le o_{i_1}, \tag{2e}$$

$$\forall i_1: \; \sum_{u\in\mathcal{U}}\sum_{m_1\in\mathcal{M}_1} a^u_{i_1,m_1,s}\cdot f^u_{i_1,m_1} \le x_{i_1}K_1, \tag{2f}$$

$$\forall m_1: \; \sum_{u\in\mathcal{U}} a^u_{0,m_1,s} \le \boldsymbol{\pi}_{m_1}, \tag{2g}$$

$$\forall i_1,m_1: \; \sum_{u\in\mathcal{U}} a^u_{i_1,m_1,s} \le \overline{\boldsymbol{\pi}}_{m_1}, \tag{2h}$$

$$\forall u: \; \sum_{m_1\in\mathcal{M}_1} a^u_{0,m_1,s}\cdot f^u_{0,m_1} + \sum_{i_1}^{N_{RRH}}\sum_{m_1\in\mathcal{M}_1} a^u_{i_1,m_1,s}\cdot f^u_{i_1,m_1} \le r^u_s. \tag{2i}$$

In P. 2, the objective (2a) is to maximize profit. The profit increases in the number of bits carried in downlink for subscribers, but decreases in the power consumption. The constant α is the conversion parameter from the number of serviced bits to what users has to pay (i.e., profit to be given to MNO_1). The $a^u_{i_1,m_1,s} \in [0,1]$ is the fraction of time [8], [39], [40] that UE u makes an association with an MBS (if $i_1 = 0$) or RRH (if $i_1 \neq 0$) and receives service on channel m_1. The $f^u_{i_1,m_1}$ is the data rate at which UE u receives data service from an MBS (if $i_1 = 0$) or RRH (if $i_1 \neq 0$) on channel m_1. The θ is a conversion parameter from the amount of power consumption to the usage fee. The τ^M (or τ^R) is the per-channel power of MBS (or RRH), and $H_s(y,\mathbf{A}_s;\xi_s)$ is the profit from `Stage-3`. Please note that `Stage-3` is

for a recourse action and is deterministic once a scenario (or network snapshot) becomes known at the beginning of each time slot. Thus, $H_s(\cdot)$ is not an expected profit.

Each UE is charged for its data usage by the first and the third term in (2a) depending on which type of BS it is associated with. The term $f^u_{i_1,m_1}$ is the channel capacity, defined as:

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

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

For each UE, the aggregate amount of time to use MNO_1's BSs cannot exceed a unit time by (2b). The constraints (2c) and (2d) indicate that each UE can access MBS and RRH, respectively, only through the channels available to the corresponding BS. Please note that in (2d), $v^u_{i_1,s} \in \mathbf{V}_s$ is a binary indicator, telling whether UE u is within the coverage of RRH i_1 or not, denoted by 1 or 0, respectively. However, we do not need it in (2c) since an MBS is assumed to be always accessible. As a reminder, the channel partition vector for MBS and RRH are $\boldsymbol{\pi}$ and $\overline{\boldsymbol{\pi}}$, respectively. The number of UEs that an RRH can handle at the same time is limited by o_{i_1} (2e),[8] while that of an MBS is assumed to be large. The fronthaul link between each RRH and the BBU pool has a limited capacity, meaning that the aggregate amount of downlink traffic carried over the link is limited (2f). To be specific, the right-hand side of (2f) is upper-bounded by the fronthaul capacity c_{i_1} in (1e). The constraints in (2g) and (2h) are to guarantee that an MBS and RRH, respectively, cannot use a channel for more than a unit time, if accessible. The total number of bits that a UE receives in downlink cannot exceed its demand by (2i).

`Stage-3` is for a recourse action. For those UEs whose service demand is not fully satisfied, they are allowed to use ON's resource through APs by offloading their traffic to ON. This is the case when the entire or a part of the MNO_1's network is saturated, and thus MNO_1 cannot meet the demand of all UEs. The problem formulation is given below (called P. 3).

$$\begin{aligned} &H_s(y, \mathbf{A}_s; \xi_s) \\ &:= \max_{\mathbf{B}_s, \widetilde{\mathbf{r}}_s} . \ \alpha\Big(\sum_{i_2 \in \mathcal{I}_2} \sum_{u \in \mathcal{U}} \sum_{m_2 \in \mathcal{M}_2} b^u_{i_2,m_2,s} \cdot g^u_{i_2,m_2}\Big) \\ &\qquad - \beta\Big(\sum_{i_2 \in \mathcal{I}_2} \sum_{u \in \mathcal{U}} \sum_{m_2 \in \mathcal{M}_2} b^u_{i_2,m_2,s} \cdot g^u_{i_2,m_2}\Big) \\ &\qquad - \gamma \sum_{u \in \mathcal{U}} \widetilde{r}^u_s \end{aligned} \tag{3a}$$

subject to

$$\forall u : \sum_{i_2 \in \mathcal{I}_2} \sum_{m_2 \in \mathcal{M}_2} b^u_{i_2,m_2,s} \le 1, \tag{3b}$$

$$\forall u, i_2, m_2 : \ 0 \le b^u_{i_2,m_2,s} \le w^u_{i_2,s}, \tag{3c}$$

$$\forall i_2 : \sum_{u \in \mathcal{U}} \sum_{m_2 \in \mathcal{M}_2} b^u_{i_2,m_2,s} \le d_{i_2}, \tag{3d}$$

$$\forall i_2 : \sum_{u \in \mathcal{U}} \sum_{m_2 \in \mathcal{M}_2} b^u_{i_2,m_2,s} \cdot g^u_{i_2,m_2} \le z_{i_2}, \tag{3e}$$

$$\forall u, i_2 : \sum_{m_2 \in \mathcal{M}_2} b^u_{i_2,m_2,s} \le 1, \tag{3f}$$

$$\begin{aligned} \forall u : &\sum_{m_1 \in \mathcal{M}_2} a^u_{0,m_1,s} \cdot f^u_{0,m_1} \\ &+ \sum_{i_1 \in \mathcal{I}_1} \sum_{m_1 \in \mathcal{M}_1} a^u_{i_1,m_1,s} \cdot f^u_{i_1,m_1} \\ &+ \sum_{i_2 \in \mathcal{I}_2} \sum_{m_2 \in \mathcal{M}_2} b^u_{i_2,m_2,s} \cdot g^u_{i_2,m_2} \\ &+ \widetilde{r}^u_s = r^u_s \end{aligned} \tag{3g}$$

$$\forall u : \ \widetilde{r}^u_s \ge 0, \tag{3h}$$

$$\begin{aligned} &\sum_{i_2 \in \mathcal{M}_2} \sum_{u \in \mathcal{U}} \sum_{m_2 \in \mathcal{M}_2} b^u_{i_2,m_2,s} \cdot g^u_{i_2,m_2} \\ &\le y \cdot K_2. \end{aligned} \tag{3i}$$

`Stage-3` has a similar problem structure to `Stage-2`. However, in this stage, MNO_1 allows its subscribers to use ON's networking resource in order to minimize service outage; otherwise, MNO_1 has to pay the penalty for unmet demand. To do so, MNO_1 pays the necessary costs in both `Stage-1` and `Stage-3`. The cost that MNO_1 is charged at `Stage-1` is to reserve ON's resource (i.e., reservation fee), while the cost at `Stage-3` is for the actual usage of it (i.e., offloading cost).

The objective (3a) is to maximize the profit, which is a function of the serviced bits via ON and the unmet demand. The parameter β is an offloading cost to ON (i.e., cost for using ON's networking resource) and γ is a penalty for unmet demand. Given the two parameters, the three terms in (3a) correspond to the profit from serviced bits, the offloading cost and the penalty for unmet demand in sequence. The total amount of time to use ON's resource cannot exceed the limit of 1 by (3b), where $b^u_{i_2,m_2,s} \in \mathbf{B}_s$ is a fraction of time UE u associates with AP i_2 to receive service over channel m_2 in scenario s. A UE can access an AP only when it is within the coverage of the AP by (3c), i.e., when $w^u_{i_2,s} = 1$. Each AP has a maximum number of UEs to which it can provide service simultaneously by (3d) with a limited bandwidth capacity (3e). Each channel cannot be occupied/used for more than a unit time by (3f).

For each UE, the sum of received service and unmet demand equals its service demand (3g). In the `Stage-3` formulation, the service demand requirement is expressed by an equality constraint by using a slack variable $\widetilde{r}^u_s$ as in

[8] Although the association capacity can be determined by other various factors such as the number of channels available to MNO_1, we assume it is dominated by the processing capacity of RRHs which is fixed during manufacturing. Please note that to make RRHs low-cost and light-weight, their hardware complexity is significantly reduced, leaving only simple functions in RRHs [3].

(3g). The amount of unmet demand for each UE is denoted by $\widetilde{r}_s^u$ which is non-negative by (3h). Please note that in this work the unmet demand is defined by the demand the primary MNO failed to satisfy even after offloading traffic to ON. Thus, it is computed at the last stage, `Stage-3`. The (3i) indicates that the aggregate amount of traffic offloaded to ON should not exceed the limit determined by the contract made in `Stage-1`. The conversion parameter K_2 is multiplied to y which is the fraction of the computing resource reserved at ON. That is, $y \cdot K_2$ indicates the maximum number of offloaded bits that can be processed by ON in a unit time.

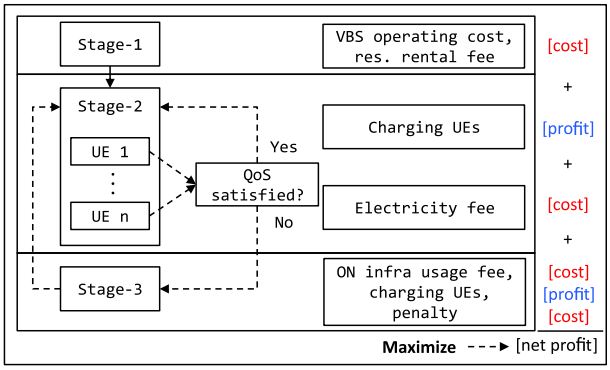

FIGURE 5. The proposed 3-stage profit and cost model computes the net profit which is the sum of all positive profit colored in blue and negative cost values colored in red.

The Fig. 5 summarizes the 3-stage profit and cost model proposed in this study, which is one of the major contributions. Depending on the decisions to make in each stage, MNO_1 pays fee/cost and/or makes a profit. In `Stage-1`, the MNO_1 makes a long-term decisions on some its networking resources without providing any service to users. Thus, there is no profit to make. In `Stage-2` and `Stage-3`, on the other hand, users receive data service from MBS/RRH or AP, respectively, and from which MNO_1 makes profit. The MNO_1 pays for the power consumed in `Stage-2`. Besides, upon the use of the ON's networking resource to offload the users' traffic, MNO_1 pays for it to compensate for the ON's operational cost in `Stage-3`. Lastly, if there is any unmet demand, MNO_1 pays a penalty in proportion to the total amount in `Stage-3`.

By combining the three per-stage problems, P. 1–P. 3, we can get the complete problem formulation in SP that optimizes the comprehensive H-CRAN resources to maximize the total profit. Although each stage optimizes a subset of H-CRAN resources, by combining them together, decisions made in each stage affect the rest and thereby yielding the optimal solution. Despite of the increased computational and space complexity, such all-in-one approach is significantly important especially for H-CRAN which consists of computing and networking resources together. If P. 1 is solved independent of the rest stages, for example, it has an access to only partial, abstracted information from networks, such as, how many jobs or bits a VBS may process, that are relevant to BBU partitioning. However, such a limited view may lead to infeasible solutions or resource under-utilization. For example, P. 1 *alone* may conclude that VBS i_1 processing $x_{i_1} \cdot K_1$ amount of traffic is the optimal solution. However, the solution may turn out to be infeasible for the power budget at the associated RRH i_1. It happens when the channel gains between the RRH and many of its associated UEs are low; for example, cell-edge users or having signal-blocking obstacles in the middle. To provide an agreed data rate to each UE, the RRH will increase the transmit power, and in the end, the required aggregate power may exceed the RRH power budget which is very low. On the contrary, when the channel gains to associated UEs are very good, the RRH can be under-utilized. It is not an efficient solution since RRH can provide high data rate with lower power than MBS. Please note that the fronthaul capacity does not cause any issue here since it is considered in (1e).

C. DETERMINISTIC EQUIVALENT PROBLEM (DEP)

The next step is to transform the SP model into a deterministic equivalent, DEP, to be able to efficiently implement and solve it with a commonly-available computer solver. Given the distribution of the uncertain parameters, we can find the probability p_s of each scenario s in the scenario tree $\mathcal{T}$. After replacing $\mathcal{Q}(\mathbf{x}, y) = \mathbb{E}_\xi[Q_s(\mathbf{x}, y; \xi_s)]$ in (1a) with the sum of p_s multiplied by the profit to make from the corresponding scenario over $\forall s \in \mathcal{S}$, and then, by combining P. 1, P. 2 and P. 3, we get the deterministic formulation of the SP as below (called P. 4).

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

$$\text{s.t.} \quad \text{constraints in (1b)-(1e), (2b)-(2i), (3b)-(3i),} \quad \text{(4a)}$$

where *s.t.* is short for subject to. Please note that the `Stage-3` profit $H_s(\cdot)$ is already included in the `Stage-2` profit $Q_x(\cdot)$, which is why $H_s(\cdot)$ is not explicitly shown in (4a). The objective (4a) is the sum of the `Stage-1` objective value and the expectation of the remaining stages' objectives. Constraints from all stages are then followed to make the DEP equivalent to the SP formulation. The Algorithm 1 in Section IV-A solves DEP (P. 4) on each iteration as shown in line 7. Here, p_s is the probability that ξ takes on ξ_s and $\mathcal{S}$ is the index set of scenarios, where $|\mathcal{S}| = |\mathcal{T}|$.

V. EVALUATION

A. NETWORK PARAMETERS

On the network we set up for evaluation, MBS has a coverage radius of 300 meters, within which four RRHs are located. Each RRH has a radius of 100 meters. RRHs have a keep-away distance of 100 meters from the MBS tower, and RRHs are equally spaced. APs are located in a similar manner, and their locations are outside the coverage of RRHs. The coverage radius of an AP is 150 meters. Please refer to Table 3 for the network parameters used for evaluation. For ϵ and δ,

TABLE 3. Network parameters.

Parameter	Value	Parameter	Value
$\lvert\mathcal{I}_1\rvert$	4	ϵ	2δ
$\lvert\mathcal{I}_2\rvert$	4	δ	\$2.608/hour [41]
$\lvert\mathcal{M}_1\rvert$	5	K_1	300Mbps
$\lvert\mathcal{M}_2\rvert$	5	K_2	300Mbps
$\lvert\mathcal{U}\rvert$	15	α	\$20/300MB/mo [42]
$\lvert\mathcal{T}'\rvert$	30	θ	\$0.0897/watt-hour [26]
o_{i_1}	3 ($\forall i_1 \in \mathcal{I}_1$)	γ	3α
d_{i_2}	3 ($\forall i_2 \in \mathcal{I}_2$)	β	\$15/500MB [10]
z_{i_2}	30Mbps ($\forall i_1 \in \mathcal{I}_1$)	c_{i_1}	30Mbps ($\forall i_2 \in \mathcal{I}_2$)

we assume that both MNO_1 and ON use a cloud computing service to process user data, such as Amazon EC2 [41]. In particular, we have chosen the reserved instance pricing model of the general purpose instance, m4. Service demand of each user follows uniform distribution whose mean value is randomly drawn from [0, 20] Mbps. Please note that in the simulated network, the actual service demand of each UE changes over time.

We have used the channel model and parameters in [43]. The distance-dependent path loss from MBS to UE and RRH to UE is PL (dB) $= 128.1 + 37.6 \cdot \log_{10}(R)$ and PL (dB) $= 140.7 + 36.7 \cdot \log_{10}(R)$, respectively, where R is the distance in km. Log-normal random variable with standard deviation of 10 dB is used to model shadowing, and a normalized Rayleigh for small-scale fading effect. The channel bandwidth is 1.25 MHz, and the total transmission power for MBS and RRH is 20 W (43 dBm) and 100 mW (20 dBm), respectively. The noise power density is -174 dBm/Hz [27]. Each slot lasts for 1 ms, which equals the transmission time interval defined in [44]. At the beginning, UEs are randomly located. Given the location of UEs, the accessibility between UE and RRH as well as UE and AP are determined periodically. The random waypoint model is used to model the mobility of mobile users.

B. SOLUTION PROCEDURE

There are uncountably many possible realizations for the service demand $\mathbf{r}$, which follows the continuous uniform distribution. Thus, the DEP problem (P. 4) is intractable due to the infinite number of scenarios to take into account to solve the problem. Thus, we apply the discretization method to $\mathbf{r}$ as aforementioned in Section III-D. In particular, we approximate each UE's demand into three discrete values to indicate low, mean and high demand. At the same time, we minimize the errors between the original distribution (i.e., continuous uniform random) and the 3-point approximation for each user; please refer to Appendix for detail. Due to the vast size of the data set or the number of scenarios in $\mathcal{T}$, however, it is still not efficient or even possible to solve P. 4 with the complete scenario tree. To be specific, for a UE u, there are $|\mathcal{I}_1| + 1$ different realizations with respect to which RRH the UE u can access, where the added one represents the case when UE u has no access to any. In the same manner, UE u has $|\mathcal{I}_2|+1$ realizations for the UE-AP accessibility. UE u has three realizations with respect to service demand. Given that there are $|\mathcal{U}|$ UEs on the network, the complete scenario tree is composed of $((|\mathcal{I}_1|+1)(|\mathcal{I}_2|+1)\times 3)^{|\mathcal{U}|}$ scenarios. For the given parameters in Table 3, there are 1.34e+28 scenarios, which makes the problem intractable in terms of time and space complexity.

In this regard, we apply the sample average approximation (SAA) method [32] to reduce the problem size or, in other words, the number of scenarios to be solved. That is, instead of solving the DEP (P. 4) with the complete scenario tree $\mathcal{T}$, a *reduced* scenario tree $\mathcal{T}'$ is used, which is constructed by randomly sampling a subset of scenarios from $\mathcal{T}$, where $|\mathcal{T}'| < |\mathcal{T}|$. Let $\mathcal{S}'$ be the scenario index set corresponding to $\mathcal{T}'$. Then, the expected `Stage-2` profit in P. 1 can be replaced by a Monte Carlo estimate $\mathcal{Q}(\mathbf{x}, y)' = \sum_{s\in\mathcal{S}'} Q_s(\mathbf{x}, y; \xi_s)/|\mathcal{S}'|$, where each scenario in $\mathcal{T}'$ is equally likely and $\xi_s \in \mathcal{T}'$. As a result, we can rewrite (4a) as follows after replacing $\mathcal{S}$ and p_s with $\mathcal{S}'$ and $1/|\mathcal{S}'|$, respectively. The resulting problem can be solved in a finite amount of time.

$$-\delta \sum_{i_1\in\mathcal{I}_1} x_{i_1} - \epsilon \cdot y + \frac{1}{|\mathcal{S}'|} \sum_{s\in\mathcal{S}'} [Q_s(\mathbf{x}, y; \xi_s)]$$

Due to the use of reduced, randomly sampled scenarios, optimal solutions vary with respect to the chosen scenarios to construct $\mathcal{T}'$. In this regard, we will show both stability and quality of solution at the end of this section. We have implemented and evaluated the proposed method as well as the ones for comparison on Matlab and CVX[9] [45]. We have constructed a scenario tree with 30 scenarios (i.e., $|\mathcal{T}'| = 30$) chosen at uniform random for Sections V-C and V-D. In Section V-F, we construct three more scenario trees in the same manner in order to study the solution stability and quality. For each simulation, we have simulated the network for 1,000 seconds. Please note that the reduced scenario tree is used to make the long-term decisions that do not change frequently, i.e., channel partitioning, BBU pool partitioning and the amount of resource to borrow from ON. Afterwards, we run the simulation with randomly generated network parameters with the long-term decisions fixed.

C. EFFECT OF UNCERTAINTIES

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

- **Constant Mobility, Constant Demand (CMCD)** which assumes constant UEs' mobility and constant service demand,[10]

[9] CVX is a Matlab-based convex modeling framework for disciplined convex programming. For further information, please refer to [45] for CVX or [46] for disciplined convex programming.

[10] In the context of SP, this type of problem is called expected value (EV) problem since uncertain parameters are replaced by their corresponding expected values which are constant.

- **Constant Mobility (CM)** which assumes constant UEs' mobility, while considering the uncertainty in demand,
- **Constant Demand (CD)** which assumes constant demand, while considering the uncertainty in UEs' mobility, and
- **proposed** which considers the uncertainties in both mobility and demand.

Please note that *constant* implies the expected value of the corresponding uncertain parameter. For example, constant demand algorithm uses the expectation of the demand which is constant when solving the optimization problem instead of considering the uncertainty in demand.

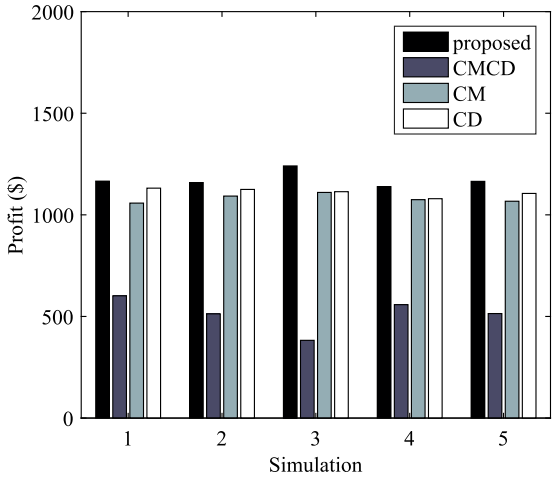

FIGURE 6. Comparison of the achieved optimal profit with respect to different ways of handling uncertainties over five runs of simulation.

Fig. 6 shows the optimal profit achieved by the four methods for five runs of simulation. Clearly, CMCD yields the least profit. This is because CMCD ignores the uncertainties in both mobility and demand by assuming constant, expected values. As a result, CMCD suffers from the under-provisioning problem; i.e., the reserved resource is not enough to handle the actual demand. Insufficient network resource reserved by CMCD soon results in the network saturation, causing users to experience service outage to a large degree as shown in Fig. 7. On the other hand, the proposed method, CM and CD consider uncertainties on the network. Such an awareness makes those methods reserve more resource than CMCD, and thus, the three methods are likely to suffer less from the under-provisioning problem. Among the three that partially or completely consider the uncertainties, the proposed method achieves the highest profit, implying the importance of the comprehensive consideration on the unpredictability. In other words, since CM and CD are unaware of the uncertainty in mobility and service demand, respectively, they have a narrower view of what might happen than the proposed method has.

As it can be clearly seen in Fig. 6, considering at least one uncertainty yields much larger profit than CMCD. In addition, the performance of CD is no less than that of CM, approaching the performance of the proposed method. From this observation, we can conclude that the uncertainties in mobility has a larger effect on the profit than the varying service demand. When there are short-range BSs are deployed and used, the connectivity between such BSs and UEs is much important. This is because compared to MBSs the short-range RRHs can provide high throughput to users with much low power by taking advantage of the short distance to users. The UE-RRH connectivity varies significantly over time due to the short service range of RRHs and the users' mobility, and thus CD outperforms CM.

In addition, CD performs not much less than the proposed method, for example, in scenario 2. It may give such an impression that considering the uncertainty in users' demand in addition to that in mobility does not provide much benefit. However, please note that the Fig. 6 shows the achieved profit for 1,000 seconds of simulation, which is much shorter than the lifetime of a cellular network standard. The seemingly-small profit difference in Fig. 6 will become much larger in the long run, and such margin can only be beaten by being aware of both uncertainties at the same time.

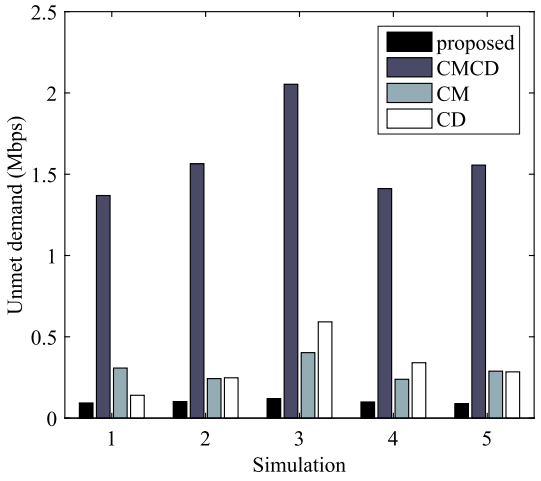

FIGURE 7. Comparison of the average unmet demand per user with respect to different ways of handling uncertainties over five runs of simulation.

The Fig. 7 depicts the average amount of unmet demand per user over five runs of simulation. It implies whether each method suffers from the resource under-provisioning problem or not. CMCD does not consider any of the network uncertainties, and thus, it optimizes the networking resource only based on the constant, expected values for users' mobility and service demand. Due to this shortsighted view over the network, CMCD reserves the least amount of resource and fulfills the constant service request coming from the stationary users. As a result, CMCD suffers most from the service outage. On the other hand, the rest three schedule more resource in order for them to be ready for many different scenarios that are likely to occur.

Although the three methods consider the uncertainties, all of them have non-zero unmet demand. This shows a tradeoff between the penalty for unmet demand and the

over-provisioning cost. Reserving too much resource on both BBU pool and ON may result in a zero unmet demand at all times. However, such an over-provisioning approach can leave a significant portion of the resource unused. That is, compared to the expected profit, the initial investment is too large. In this regard, what Fig. 7 shows is that instead of having all demand perfectly satisfied, allowing a negligible amount of QoS degradation gives a higher profit gain overall. Again, the proposed method results in the least amount of unmet demand among the four, showing that it is beneficial to be aware of uncertainties as much as possible.

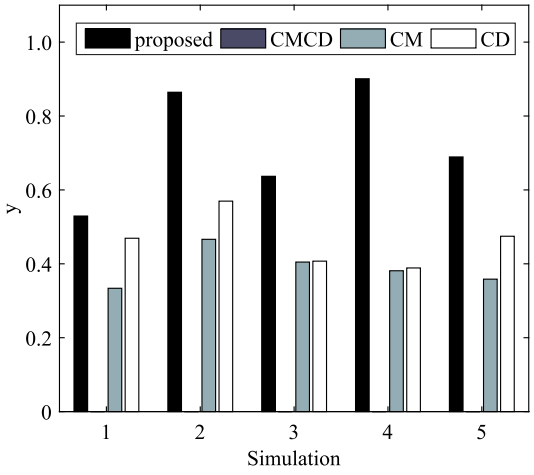

FIGURE 8. Comparison of the fraction (y) of the resource that MNO_1 borrows from ON at `Stage-1` with respect to different ways of handling uncertainties over five runs of simulation.

Continuing the discussion on the mean unmet demand, Fig. 8 shows why CMCD results in large unmet demand. The y axis in Fig. 8 indicates how much resource does MNO_1 borrow from ON. In the case of the proposed method, for having more knowledge on uncertainties, MNO_1 borrows more resource from ON. However, that is not the case to CMCD. Since CMCD is informed only of one snapshot of the network as to users' location and service demand, CMCD does not borrow any resource from ON, resulting in a large unmet demand as it can be seen in the previous Fig. 7.

Next, we have measured the energy consumption and energy efficiency. First, the Fig. 9 shows the total amount of power consumption of the four different schemes over five runs of simulation. Although the proposed scheme spens the least amount of power during operation, the difference among the four schemes is not significant. However, considering the profit in Fig. 7 together, the energy efficiency differs to a large degree between different schemes as shown in Fig. 10.

Instead of considering only the power consumption, Fig. 10 takes both profit and power consumption into account and depicts the amount of power consumed to make a unit profit (i.e., a dollar). The inverse of the reported values in the figure becomes an energy-efficiency metric, which we omitted

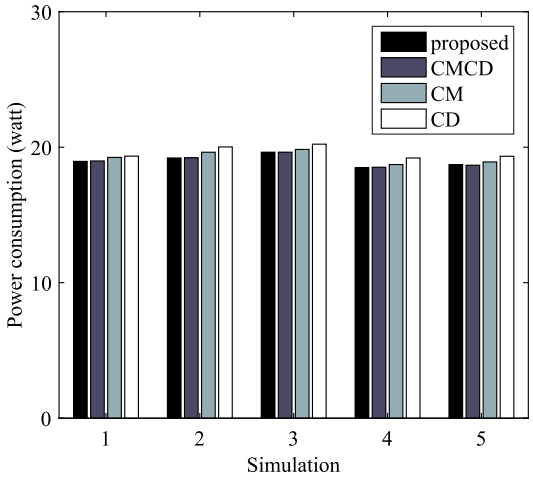

FIGURE 9. Total power consumption with respect to different ways of handling uncertainties over five runs of simulation.

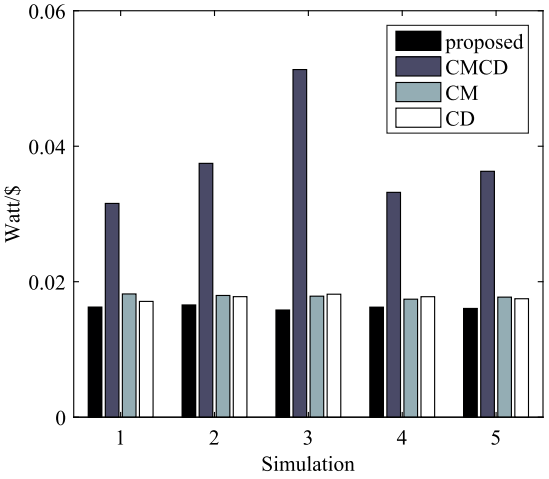

FIGURE 10. Comparison of the power consumed to make a unit profit ($) with respect to different ways of handling uncertainties over five runs of simulation. Please note that the inverse of each value in this figure represents the proposed energy-efficiency metric, i.e., ($-per-Watt).

since it can be simply computed. Clearly, CMCD consumes much more power compared to the rest three methods to make a unit profit, resulting in a low energy efficiency. Given that CMCD has suffered most from the unmet demand, it has made the least profit from providing service to users, while paying the most penalty for unmet demand. On the other hand, CM, CD and the proposed methods have yielded a relatively small amount of unmet demand while consuming a similar amount of power to CMCD. As a result, they can achieve a much higher profit and the energy efficiency than CMCD. Still, the proposed scheme outperforms the rest since it has recorded the least power consumption and unmet demand among the four methods, while achieving the most profit.

In what follows, we compare the per-stage profit of the four schemes, and study how they contribute to the total profit.

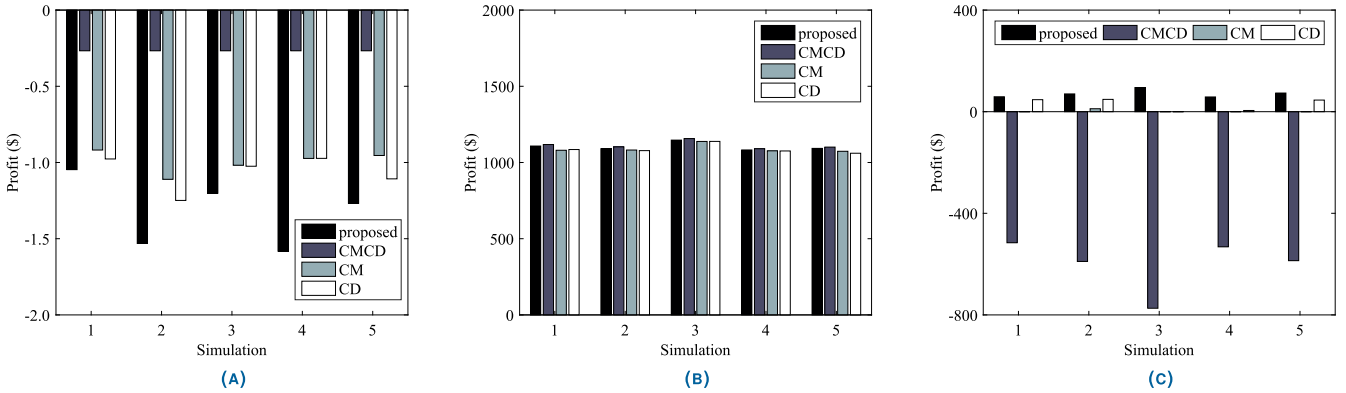

FIGURE 11. Comparison of the per-stage profit with respect to different ways of handling uncertainties over five runs of simulation. (a) Stage-1 profit. (b) Stage-2 profit. (c) Stage-3 profit.

Fig. 11 shows the `Stage-1` profit of the four schemes. The profit in this stage cannot be positive since MNO_1 does not provide any service to users. Instead, it pays for both operating its resource (i.e., BBU pool) and borrowing resource from ON. The expense for operating BBU is not much different across the four scheme, and thus, the difference in the `Stage-1` profit mainly comes from the amount of resource borrowed from ON to prepare for possible network saturation. The proposed scheme, CM and CD are fully or partially aware of the uncertainties in the network and thus borrow a certain amount of resource from ON. However, CMCD is not, and thus, it concludes that the users' demand can be satisfied solely by using MNO_1's resource. As a result, CMCD spends the least in `Stage-1`.

The effects of the decision made in `Stage-1` does not become evident in `Stage-2` yet. The Fig. 11b shows the `Stage-2` profit, showing that the difference in the `Stage-2` profit is trivial among the four schemes. In fact, CMCD achieves the most profit, implying that CMCD is heavily using the MNO_1's resource to fulfill the users' demand. It would be an optimal decision if there was no uncertainty on the network without any unmet demand, but it is not the case as shown in the following Fig. 11c.

Fig. 11c shows the `Stage-3` profit. In `Stage-3`, some traffic is offloaded to ON, and the penalty to the resulting unmet demand is measured and charged. Since there is no such scheme that completely fulfills the service demand, all schemes pay the price for the unmet demand which affects the overall revenue. However, only CMCD has yielded a negative profit in `Stage-3` since it did not reserve any resource at ON while having much unmet demand. Thus, in the case of CMCD, all unmet demand at the end of `Stage-2` equals that of `Stage-3`. However, the rest three have resulted in nonnegative profit in `Stage-3`. That is, although each of the three schemes has a strongly positive amount of penalty for the unmet demand, it is offset by the profit from providing service to users through ON's network. In addition, the proposed method has successfully made a strongly positive profit in `Stage-3` over all runs of simulation, showing the advantage of being well aware of both network uncertainties. In sum, being well aware of the network uncertainties can yield a more profitable and energy-efficient solution compared to others that are not.

D. EFFECT OF DIFFERENT OPERATION RULES

In this section, we study the effects of different *operation rules* on the performance. The proposed method allows a negligible amount of QoS degradation at the expense of penalty, which is done by introducing a nonnegative slack variable $\widetilde{r}_s^u$ as in (3g) and (3h). Also, by allowing traffic offloading to another MNO, the proposed method minimizes unmet demand while maximizing profit. In order to study the effect of such operation rules, we compare the proposed method to other models with different operation rules. Please note that the methods to be used for comparison in this section are aware of the network uncertainties as much as the proposed method is. The models to be considered in this section are:

- **Perfect Service Demand (PSD)** which has to perfectly satisfy service demand, and does not allow any unmet demand.
- **No Offloading (NoOFL)** which does not allow any traffic offloading to another MNO.
- **proposed** which allows both service outage (with penalty) and traffic offloading.

We evaluated the three different rules with respect to the same evaluation criterion as in Section V-C. We have carried out five runs of simulation, and the averaged results are summarized in Table 4.

As it can be seen in the table, PSD resulted in an infeasible solution at all runs of simulation, and thus it could not provide any service to users. Due to the mobility and varying service demand of users, MNO_1 cannot always fully satisfy the users' demand in the given network setting. This implies that MNOs

TABLE 4. Comparison of the mean performance out of five runs of simulation.

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

should not put a PSD-like rule into effect since it might put the entire system into an inoperable state when there is no feasible solution.

Also, Table 4 shows that the proposed method much outperforms NoOFL. The achieved profit of the proposed method is almost twice that of NoOFL. The main reason for such a significant difference in profit is the penalty to pay for unmet demand in `Stage-3`. In contrast to the proposed method which made a positive profit in `Stage-3`, NoOFL resulted in a large expense in the same stage mainly due to the large amount of unmet demand. Considering that both methods used similar amount of power during operation, the proposed method significantly outperforms NoOFL in terms of the energy efficiency as well. Since NoOFL does not allow any offloading to ON, its optimal solution and the resulting performance are similar to that of CMCD. It is noteworthy that CMCD and NoOFL are totally different in terms of the awareness of the uncertainties. However, due to the fact that both do not or cannot leverage any resource from ON, the performance of one becomes similar to that of the other. This result proves the significance of allowing traffic offloading which is as important as considering the uncertainties in the network.

In sum, a strict operation rule that compels an MNO to fully satisfy the users' service demand should not be considered in practice since it can cause the entire network to be unavailable. Also, being capable of traffic offloading significantly increases the profit and energy efficiency, since a network can make use of more networking resource than it actually owns. In other words, sharing the network resource among different MNOs is an effective way of enhancing the network capacity without purchasing additional bandwidth or RRHs.

E. TIME COMPLEXITY: RUNTIME

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

F. STABILITY AND QUALITY OF SOLUTION

As aforementioned, we have used a sampling method as part of the solution procedure to reduce the problem size so that a computer solver can solve the optimization problem in a finite amount of time. To be specific, out of all possible combinations of the uncertainty parameters describing the connectivity and service demand, we draw only a subset at uniform random to construct a reduced scenario tree $\mathcal{T}'$. Thus, the optimal solutions from different $\mathcal{T}'$ can vary depending on the scenarios included therein. In this regard, it is important to evaluate the stability and quality of the solution. In this section, we first show that the solutions from different scenario trees have in-sample and (weaker) out-of-sample stability, meaning that the solution does not depend much on the construction of scenario trees. Also, by evaluating the solution quality, we show that the upper bound of the error caused by SAA is small. Interested readers can refer to [33] for an in-depth discussion on the stability and quality of solution for SP.

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

1) IN-SAMPLE STABILITY

The solution procedure has an in-sample stability if the following is satisfied [33]: $\phi(\hat{\mathbf{x}}_i; \mathcal{T}_i) \approx \phi(\hat{\mathbf{x}}_j; \mathcal{T}_j)$, which happens when the selection of scenario tree does not affect the optimal objective value much. Thus, if we have an in-sample stability, we can choose any scenario tree when solving a stochastic program, and the corresponding optimal solution will be comparable to the ones from other scenario trees. We have used the Jain's fairness index [47], $\mathcal{J}(\rho_1, \rho_2, \cdots, \rho_n) = \frac{(\sum_{i=1}^{n} \rho_i)^2}{n \times \sum_{i=1}^{n} \rho_i^2}$, to check how much are the objective values from different scenario trees close to each other. Out of four scenario trees we have constructed, we have a fairness index of $\mathcal{J}(\{\rho_{ii} | \forall i = 1, 2, 3, 4\}) = 0.9999$ which is close to 1. That is, the optimal objective values from different scenario trees are close to each other. It implies that no matter how each scenario tree is configured, the optimal value (i.e., optimal

TABLE 5. Results for the weaker out-of-sample stability.

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

profit) from such a scenario tree will not much differ from the others from different scenario trees.

2) WEAKER OUT-OF-SAMPLE STABILITY

The weaker out-of-sample stability checks if the solution from a scenario tree is still a good solution to other scenario trees as well. We have a weaker out-of-sample stability, if the following holds [33]: $\phi(\hat{\mathbf{x}}_i;\mathcal{T}_j) \approx \phi(\hat{\mathbf{x}}_j;\mathcal{T}_i)$. The Table 5 shows the weaker out-of-sample stability test results. The $(i,j)^{\text{th}}$ entry in the table is the results of $|\rho_{ij}-\rho_{ji}|$. The mean of the values above the diagonal is 30.6819 which is merely 2.64% of the mean of ρ_{ij} for all i,j where $i\neq j$. Therefore, the solution procedure has a weaker out-of-sample stability.

3) SOLUTION QUALITY

Since the solution procedure has both in- and out-of-sample stability, we can study the solution quality. By definition [33], the quality of a given solution $\hat{\mathbf{x}}^i$ corresponds to the optimality gap which is defined as $\text{err}(\hat{\mathbf{x}}_i) = \max_{\mathbf{x}}\phi(\mathbf{x};\xi) - \phi(\hat{\mathbf{x}}_i;\xi)$. However, it is impossible to find the gap because of the vast size of the original problem that cannot be solved efficiently. Instead, we calculate a statistical estimate of the error of a particular solution $\hat{\mathbf{x}}_i$, which is: $\text{err}(\hat{\mathbf{x}}_i) \lesssim \frac{1}{n}\sum_{j=1}^{n}[\max_{\mathbf{x}}\phi(\mathbf{x};\mathcal{T}_j) - \phi(\hat{\mathbf{x}}_i;\mathcal{T}_j)]$, where $n=4$ and $\leq$ can be used instead of $\lesssim$ as $n\to\infty$. The Table 6 shows the stochastic upper bound on the error for each solution. Compared to the objective values which are greater than or equal to 1114.3297 in any cases, the error is small, implying a high quality of the solutions.

TABLE 6. Stochastic upper bound on the error for each solution.

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

VI. CONCLUSION

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

APPENDIX. THREE-POINT APPROXIMATION

In order to discretize a uniform random variable into three discrete points, we have used the method proposed in [48]. Let G be the distribution of continuous Uniform(a,b) with mean m. Also, let G^d be the distribution of the discretized G. By using the mass transportation problem framework [48], the objective of the discretization is to minimize the distance d between the two distributions which is defined as:

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

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

Then, the discretization (or 3-point approximation) problem can be written as: $\min_{\mathbf{z}}.\ d(G,G^d) = \int_{\frac{-\infty+z_1}{2}}^{\frac{z_1+z_2}{2}} |u-z_1|\ dG(u) + \int_{\frac{z_1+z_2}{2}}^{\frac{z_2+z_3}{2}} |u-z_2|\ dG(u) + \int_{\frac{z_2+z_3}{2}}^{\frac{z_3+\infty}{2}} |u-z_3|\ dG(u)$.

After manipulating the right-hand side, we can transform it into a simple form: $\mathbf{z}^T\mathbf{Q}\mathbf{z}+f(\mathbf{z})$, where $\mathbf{z}=[z_1,z_2,z_3]^T, f(\mathbf{z})$ is an affine function of $\mathbf{z}$, and $\mathbf{Q}$ is symmetric and positive definite. Thus, there exists a unique optimal solution to the following quadratic problem (called P. 5).

$$\begin{aligned}
\min_{\mathbf{z}}.\quad & \mathbf{z}^T\mathbf{Q}\mathbf{z}+f(\mathbf{z}) && \text{(5a)}\\
\text{subject to}\quad & \mathbf{z}\succeq\mathbf{0}, && \text{(5b)}\\
& a\leq z_1\leq z_2\leq z_3\leq b, && \text{(5c)}\\
& z_2=m, && \text{(5d)}
\end{aligned}$$

where $\succeq$ is an element-wise greater than or equal to operator. The solution is non-negative by (5b), and also is bounded by the lower and upper bound, i.e., a and b, respectively by (5c). Finally, the mean value should remain the same by (5b) as that of the original distribution. The optimal solution, $\mathbf{z}$, is a vector of three equally likely discrete values that are approximations of the original uniform distribution with the same mean value.

REFERENCES

[1] Y. Lin, L. Shao, Z. Zhu, Q. Wang, and R. K. Sabhikhi, "Wireless network cloud: Architecture and system requirements," *IBM J. Res. Develop.*, vol. 54, no. 1, pp. 4:1–4:12, Jan./Feb. 2010.

[2] *C-RAN: The Road Towards Green RAN, Version 2.5*, China Mobile Research Institute, Beijing, China, Oct. 2011.

[3] J. Wu, Z. Zhang, Y. Hong, and Y. Wen, "Cloud radio access network (C-RAN): A primer," *IEEE Netw.*, vol. 29, no. 1, pp. 35–41, Jan. 2015.
[4] M. Peng, Y. Sun, X. Li, Z. Mao, and C. Wang, "Recent advances in cloud radio access networks: System architectures, key techniques, and open issues," *IEEE Commun. Surveys Tuts.*, vol. 18, no. 3, pp. 2282–2308, 3rd Quart., 2016.
[5] M. Peng, Y. Li, J. Jiang, J. Li, and C. Wang, "Heterogeneous cloud radio access networks: A new perspective for enhancing spectral and energy efficiencies," *IEEE Wireless Commun.*, vol. 21, no. 6, pp. 126–135, Dec. 2014.
[6] A. De Domenico, E. C. Strinati, and A. Capone, "Enabling green cellular networks: A survey and outlook," *Comput. Commun.*, vol. 37, pp. 5–24, Jan. 2014.
[7] D. Pompili, A. Hajisami, and H. Viswanathan, "Dynamic provisioning and allocation in cloud radio access networks (C-RANs)," *Ad Hoc Netw.*, vol. 30, pp. 128–143, Jul. 2015.
[8] A. Abdelnasser, E. Hossain, and D. I. Kim, "Tier-aware resource allocation in OFDMA macrocell-small cell networks," *IEEE Trans. Commun.*, vol. 63, no. 3, pp. 695–710, Mar. 2015.
[9] S. Luo, R. Zhang, and T. J. Lim, "Downlink and uplink energy minimization through user association and beamforming in C-RAN," *IEEE Trans. Wireless Commun.*, vol. 14, no. 1, pp. 494–508, Jan. 2015.
[10] C.-C. Hsu, J. M. Chang, and Y.-W. Chen, "Joint optimization for cell configuration and offloading in heterogeneous networks," in *Proc. IEEE Int. Conf. Comput. Commun. (INFOCOM)*, San Francisco, CA, USA, Apr. 2016, pp. 1–9.
[11] X. Kang, Y.-K. Chia, S. Sun, and H. F. Chong, "Mobile data offloading through a third-party WiFi access point: An operator's perspective," *IEEE Trans. Wireless Commun.*, vol. 13, no. 10, pp. 5340–5351, Oct. 2014.
[12] S. Li, J. Huang, and S.-Y. R. Li, "Revenue maximization for communication networks with usage-based pricing," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Honolulu, HI, USA, Nov./Dec. 2009, pp. 1–6.
[13] X. Wang, K. Wang, S. Wu, S. Di, K. Yang, and H. Jin, "Dynamic resource scheduling in cloud radio access network with mobile cloud computing," in *Proc. IEEE/ACM Int. Symp. Qual. Service (IWQoS)*, Beijing, China, Jun. 2016, pp. 1–6.
[14] C. Pan, H. Zhu, N. J. Gomes, and J. Wang, "Joint precoding and RRH selection for user-centric green MIMO C-RAN," *IEEE Trans. Wireless Commun.*, vol. 16, no. 5, pp. 2891–2906, May 2017.
[15] A. Younis, T. X. Tran, and D. Pompili, "Fronthaul-aware resource allocation for energy efficiency maximization in C-RANs," in *Proc. IEEE Int. Conf. Autonomic Comput. (ICAC)*, Trento, Italy, Sep. 2018, pp. 91–100.
[16] C. W. Patterson, A. B. MacKenzie, S. Glisic, B. Lorenzo, J. Röning, and L. A. DaSilva, "An economic model of subscriber offloading between mobile network operators and WLAN operators," in *Proc. Int. Symp. Modeling Optim. Mobile, Ad Hoc Wireless Netw. (WiOpt)*, Hammamet, Tunisia, May 2014, pp. 444–451.
[17] K. Zhu, E. Hossain, and D. Niyato, "Pricing, spectrum sharing, and service selection in two-tier small cell networks: A hierarchical dynamic game approach," *IEEE Trans. Mobile Comput.*, vol. 13, no. 8, pp. 1843–1856, Aug. 2014.
[18] M. Y. Lyazidi, N. Aitsaadi, and R. Langar, "Dynamic resource allocation for cloud-RAN in LTE with real-time BBU/RRH assignment," in *Proc. IEEE Int. Conf. Commun. (ICC)*, Kuala Lumpur, Malaysia, May 2016, pp. 1–6.
[19] L. Feng, W. Li, P. Yu, and X. Qiu, "An enhanced OFDM resource allocation algorithm in C-RAN based 5G public safety network," *Mobile Inf. Syst.*, vol. 2016, Jul. 2016, Art. no. 9586287. [Online]. Available: https://www.hindawi.com/journals/misy/2016/9586287/. doi: 10.1155/2016/9586287.
[20] J. Tang, W. P. Tay, and T. Q. S. Quek, "Cross-layer resource allocation with elastic service scaling in cloud radio access network," *IEEE Trans. Wireless Commun.*, vol. 14, no. 9, pp. 5068–5081, Sep. 2015.
[21] S. Gu, Z. Li, C. Wu, and H. Zhang, "Virtualized resource sharing in cloud radio access networks through truthful mechanisms," *IEEE Trans. Commun.*, vol. 65, no. 3, pp. 1105–1118, Mar. 2017.
[22] W. Zhao and S. Wang, "Traffic density-based RRH selection for power saving in C-RAN," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 12, pp. 3157–3167, Dec. 2016.
[23] Y. Cai, F. R. Yu, and S. Bu, "Cloud radio access networks (C-RAN) in mobile cloud computing systems," in *Proc. IEEE Conf. Comput. Commun. Workshops (INFOCOM WKSHPS)*, Toronto, ON, Canada, Apr./May 2014, pp. 369–374.
[24] Y. Cai, F. R. Yu, and S. Bu, "Dynamic operations of cloud radio access networks (C-RAN) for mobile cloud computing systems," *IEEE Trans. Veh. Technol.*, vol. 65, no. 3, pp. 1536–1548, Mar. 2016.
[25] L. Mashayekhy, M. M. Nejad, and D. Grosu, "Physical machine resource management in clouds: A mechanism design approach," *IEEE Trans. Cloud Comput.*, vol. 3, no. 3, pp. 247–260, Jul. 2015.
[26] S. Chaisiri, B.-S. Lee, and D. Niyato, "Optimization of resource provisioning cost in cloud computing," *IEEE Trans. Services Comput.*, vol. 5, no. 2, pp. 164–177, Apr./Jun. 2012.
[27] K. Guo, M. Sheng, J. Tang, T. Q. S. Quek, and Z. Qiu, "Exploiting hybrid clustering and computation provisioning for green C-RAN," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 12, pp. 4063–4076, Dec. 2016.
[28] J. Li, M. Peng, Y. Yu, and Z. Ding, "Energy-efficient joint congestion control and resource optimization in heterogeneous cloud radio access networks," *IEEE Trans. Veh. Technol.*, vol. 65, no. 12, pp. 9873–9887, Dec. 2016.
[29] M. Peng, Y. Yu, H. Xiang, and H. V. Poor, "Energy-efficient resource allocation optimization for multimedia heterogeneous cloud radio access networks," *IEEE Trans. Multimedia*, vol. 18, no. 5, pp. 879–892, May 2016.
[30] Q. Liu, T. Han, N. Ansari, and G. Wu, "On designing energy-efficient heterogeneous cloud radio access networks," *IEEE Trans. Green Commun. Netw.*, vol. 2, no. 3, pp. 721–734, Sep. 2018.
[31] C. Pan, H. Zhu, N. J. Gomes, and J. Wang, "Joint user selection and energy minimization for ultra-dense multi-channel C-RAN with incomplete CSI," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 8, pp. 1809–1824, Aug. 2017.
[32] J. R. Birge and F. Louveaux, *Introduction to Stochastic Programming* (Series in Operations Research and Financial Engineering). New York, NY, USA: Springer, 2011.
[33] A. J. King and S. W. Wallace, *Modeling with Stochastic Programming* (Operations Research and Financial Engineering). New York, NY, USA: Springer, 2012.
[34] M. Peng, C. Wang, V. Lau, and H. V. Poor, "Fronthaul-constrained cloud radio access networks: Insights and challenges," *IEEE Wireless Commun.*, vol. 22, no. 2, pp. 152–160, Apr. 2015.
[35] D. Pompili, A. Hajisami, and T. X. Tran, "Elastic resource utilization framework for high capacity and energy efficiency in cloud RAN," *IEEE Commun. Mag.*, vol. 54, no. 1, pp. 26–32, Jan. 2016.
[36] N. Bhushan *et al.*, "Network densification: The dominant theme for wireless evolution into 5G," *IEEE Commun. Mag.*, vol. 52, no. 2, pp. 82–89, Feb. 2014.
[37] D. Tse and P. Viswanath, *Fundamentals of Wireless Communication*. Cambridge, U.K.: Cambridge Univ. Press, 2005.
[38] MathWorks. (2018). *MATLAB and Statistics Toolbox Release 2018b*. [Online]. Available: https://www.mathworks.com
[39] Z. Shen, J. G. Andrews, and B. L. Evans, "Adaptive resource allocation in multiuser OFDM systems with proportional rate constraints," *IEEE Trans. Wireless Commun.*, vol. 4, no. 6, pp. 2726–2737, Nov. 2005.
[40] M. Tao, Y.-C. Liang, and F. Zhang, "Resource allocation for delay differentiated traffic in multiuser OFDM systems," *IEEE Trans. Wireless Commun.*, vol. 7, no. 6, pp. 2190–2201, Jun. 2008.
[41] Amazon.com. *Amazon Elastic Computing Cloud (Amazon EC2)*. Accessed: Dec. 10, 2017. [Online]. Available: https://aws.amazon.com/ec2
[42] AT&T. [Online]. Available: https://www.att.com/
[43] *Evolved Universal Terrestrial Radio Access (E-UTRA); Further Advancements for E-UTRA Physical Layer Aspects*, document TR 36.814, Release 9, 3rd Generation Partnership Project (3GPP), Mar. 2010.
[44] *Evolved Universal Terrestrial Radio Access (E-UTRA); Medium Access Control (MAC) Protocol Specification*, document TS 36.321, Release 12, Apr. 2015.
[45] CVX Research. (2018). *CVX: MATLAB Software for Disciplined Convex Programming, Version 2.1.*. [Online]. Available: http://cvxr.com/
[46] M. Grant, S. Boyd, and Y. Ye, "Disciplined convex programming," *Global Optimization*. Boston, MA, USA: Springer, 2016, pp. 155–210.
[47] R. Jain, D. Chiu, and W. Hawe, "A quantitative measure of fairness and discrimination for resource allocation in shared computer system," Digital Equipment Corp., Hudson, MA, USA, Tech. Rep. DEC-TR-301, Sep. 1984.
[48] G. C. Pflug, "Scenario tree generation for multiperiod financial optimization by optimal discretization," *Math. Program.*, vol. 89, no. 2, pp. 251–271, 2001.

TAEWOON KIM (M'18) received the B.S. degree in computer science and engineering from Pusan National University, South Korea, in 2008, the M.S. degree in information and mechatronics from the Gwangju Institute of Science and Technology, South Korea, in 2010, and the Ph.D. degree in computer engineering from Iowa State University, Ames, IA, in 2018. From 2010 to 2013, he was a Research Engineer with the Telecommunications Technology Association, South Korea. He is currently an Assistant Professor with the School of Software, Hallym University, South Korea. His research interests include network modeling, optimization, and protocol design for wireless networking systems, such as WLAN, the IoT/sensor networks, heterogeneous networks, and 5G.

J. MORRIS CHANG (SM'08) received the Ph.D. degree from North Carolina State University. His past industrial experiences include positions at Texas Instruments, the Microelectronic Center of North Carolina, and AT&T Bell Labs. He is currently a Professor with the Department of Electrical Engineering, University of South Florida. In the last five years, his research projects on cyber security have been funded by DARPA. He is leading a DARPA project under the Brandeis Program, focusing on privacy-preserving computation over the Internet. His research interests include cyber security, wireless networks, and energy-efficient computer systems. He received the University Excellence in Teaching Award from the Illinois Institute of Technology, in 1999. He is a Handling Editor of the *Microprocessors and Microsystems* and the Associate Editor-in-Chief of the *IEEE IT Professional*.