Joint Resource Allocation and Admission Control in Sliced Fog Radio Access Networks

Yuan Ai[1], Gang Qiu[2], Chenxi Liu[1,*], Yaohua Sun[1]

[1] State Key Laboratory of Network and Switching Technology, Beijing University of Posts and Telecommunications, Beijing 100876, China
[2] LTE&5G Wireless Product R&D Institute, ZTE Corporation, Shenzhen 518057, China
* The corresponding author, email: chenxi.liu@bupt.edu.cn

Abstract: Network slicing based fog radio access network (F-RAN) has emerged as a promising architecture to support various novel applications in 5G-and-beyond wireless networks. However, the co-existence of multiple network slices in F-RANs may lead to significant performance degradation due to the resource competitions among different network slices. In this paper, the downlink F-RANs with a hotspot slice and an Internet of Things (IoT) slice are considered, in which the user equipments (UEs) of different slices share the same spectrum. A novel joint resource allocation and admission control scheme is developed to maximize the number of UEs in the hotspot slice that can be supported with desired quality-of-service, while satisfying the interference constraint of the UEs in the IoT slice. Specifically, the admission control and beamforming vector optimization are performed in the hotspot slice to maximize the number of admitted UEs, while the joint sub-channel and power allocation is performed in the IoT slice to maximize the capability of the UEs in the IoT slice tolerating the interference from the hotspot slice. Numerical results show that our proposed scheme can effectively boost the number of UEs in the hotspot slice compared to the existing baselines.

Keywords: NOMA; fog radio access networks; resource allocation; admission control

Received: May 4, 2019
Revised: Jun. 8, 2020
Editor: Zhongyuan Zhao

I. INTRODUCTION

Pervasive mobile internet and the Internet of Things (IoT) are enabling many innovative applications, such as virtual/augmented reality (VR/AR), vehicle-to-vehicle communications, mobile gaming, and e-health [1]. These applications require that ubiquitous services can be provided for massive number of devices with ultra-high data rate and ultra-low latency, thus introducing significant challenges to current wireless networks of centralized cloud architecture [2]. In addition, the practical fronthaul is often capacity constrained or delay constrained, which presents a bottleneck to the system performance of cloud radio access networks. To address this issue, fog radio access networks (F-RANs) have been proposed to achieve high spectral efficiency, energy efficiency, and low latency by fully utilizing the signal processing, resource management and storage capabilities of edge devices [3].

Network slicing has been advocated as a cost-efficient solution to meet diverse applications and services in wireless network. Based on the concept of network slicing, the common

physical network is sliced into multiple virtual networks, each of which consists of a set of network functions and resources [4]. The hierarchical structure of F-RANs makes it suitable for network slicing and some works have been done to show the benefits of network slicing in F-RANs [5], [6]. In F-RANs, fog access points (FAPs) and the remote radio heads (RRHs) coexist to serve their corresponding associated user equipments (UEs). According to different service requirements, UEs can select to access FAP or RRH. In the RRH mode, the user equipments are associated with the RRHs, the RRHs forward the received radio signals to the centralized baseband unit (BBU) pool, and the BBU pool executes all the functions globally and centrally. In the FAP mode, the user equipments are associated with the FAPs. Each FAP is co-located with a cache unit, and delivers the data to the UEs by exploiting cached contents or leveraging processing at the edge of the network. Note that the UEs associated with the RRHs benefit from centralized signal processing and resource allocation, and therefore can enjoy high data rate. In fact, it has been shown that the physical layer integration technique [7] in the RRH mode can effectively improve the network performance. In addition to transmit data to individual users (i.e., broadcast/unicast services), the RRHs can integrate additional multicast services [8]. That is, the RRHs transmit a common message in such a way that all the UEs in the same group can decode it. Compared with point-to-point unicast transmission, multi-group multicast transmission provides high capacity service for common content delivery to multiple users on a same radio resource block. For instance, in the scenarios of wireless video delivery [9], it is common that multiple users are interested in the same video stream, which creates multicast groups. On the other hand, as the new designed AP, the FAP integrates not only the front radio frequency (RF) but also the physical processing functionalities of the upper layers, thus making the FAP have a sufficient computing capabilities to execute the local cooperative signal processing in the physical layer and implement the caching resource management. The FAP can provide a much shorter end-to-end latency at the application layer, since the delay in the fronthaul can be omitted due to the local function. Meanwhile, NOMA techniques can achieves significant improvement in massive connectivity by allowing multiple users to share the same sub-channel in power domain.

Although sliced F-RANs have a great potential of meeting stringent requirements of future wireless networks, how to achieve the coexistence of multiple slices in the F-RANs faces many challenges. Firstly, existing resource management solutions in multi-slice F-RANs are commonly based on the slice isolation, which allocates non-overlapping resources to different slices. As such, any change in one slice instance will have no impact on the performance of the other slice instance. However, slice isolation mitigates the interference at the cost of resource usage efficiency, thus making it less suitable for future wireless networks, where wireless radio resources will become even more scarce [10]. Against this backdrop, it is pivotal to characterize the performance of F-RANs considering interactions between different slice instances. Secondly, many research efforts have been devoted to improve the spectral efficiency and/or the energy efficiency of multi-slice F-RANs. However, the number of users that can be served with desired quality-of-service (QoS) is a pivotal design aspect of multi-slice F-RANs as well, yet receives much less attention. Thirdly, motivated by the high spectral efficiency and the high user connectivity provided by the non-orthogonal multiple access (NOMA) technique, incorporating NOMA in multi-slice F-RANs has been receiving increasing research attention. However, the existing works mainly focused on the F-RANs of slice isolation, which leaves the performance of NOMA in F-RANs with interactions between different slices an open problem.

In this paper, the performance of two-slice F-RANs is investigated, where two network slices instances for the hotspot scenario and

In this paper, a framework of joint resource allocation and admission control was proposed for sliced F-RANs, in which NOMA technique and multicast beamforming are adopted in the IoT slice instance and the hotspot slice instance, respectively.

IoT scenario are considered, respectively. NOMA technique is invoked in the IoT slice instance and multicast beamforming is incorporated in the hotspot slice instance. Taking the interactions between the hotspot slice instance and the IoT slice instance into consideration, novel algorithms are developed to *maximize the number of UEs in the hotspot slice instance that can be supported with desired QoS, while simultaneously satisfying the constraints of UEs in the IoT slice instance on the minimum rate.* The main contributions of this paper are summarized as follows.

1) The impact of the IoT slice instance on the performance of the hotspot slice instance is characterized, which enables us to develop a framework to solve the considered optimization problem. In the proposed framework, the considered optimization problem can be decoupled to a joint sub-channel assignment and power allocation problem in the IoT slice instance and an admission control problem in the hopspot slice instance.

2) The joint optimization problem in the IoT slice instance is solved to maximize the total interference threshold that the IoT UEs can tolerate from the hotspot slice instance, while satisfying the QoS constraints of the IoT UEs. Specifically, we first decouple the subchannel assignment and power allocation problem as a many-to-one matching game with peer effects and a linear programming problem, respectively. Then, we solve these two problems iteratively.

3) The admission control problem in the hotspot slice instance is solved to maximize the number of admitted hotspot UEs that can achieve their QoS targets, while satisfying the interference constraints on the IoT UEs and the power budget limitations of RRHs.

The remainder of this paper is organized as follows. Section II reviews related works. Section III presents the system model and formulate the problem. Section IV and Section V focus on the optimization problem in the IoT slice instance and the hotspot slice instance, respectively. Section VI presents the simulation results. Finally, Section VII draws the conclusion.

II. Related Work

Considering the user of slice isolation, resource management solutions in multi-slice F-RANs have been extensively investigated [6], [11], [12]. Specifically, in [11], the problem of dynamic mode selection was formulated as an evolutionary game, in which the groups of potential users' space compete with each other. With the target of minimizing the long-term system power consumption under the dynamics of edge caching states, the authors in [12] developed a deep reinforcement learning based approach for F-RANs. In [6], a hierarchical radio resource allocation architecture was proposed to assign the sub-channels to the UEs operating in different slices. However, the resource usage efficiency of slice isolation is relatively low, thus limiting its applicability in future wireless networks. To address this issue, it is indeed important to investigate the impact of resource allocation in one slice instance on the other slice instance.

When the number of the UEs is relatively large, it is very unlikely for wireless networks to support all the UEs with desired QoS. To address this issue, admission control was proposed and widely adopted to temporarily drop some UEs such that the admitted UEs can be served with satisfied QoS [13]- [16]. Using admission control, in [13], low-complexity convex approximation algorithms were proposed to maximize the number of users that can be served at their desired QoS. In [14], a holistic sparse optimization framework was developed for network power minimization and user admission in the C-RANs. In [15], a novel joint admission control and resource allocation strategy was proposed to satisfy the UEs' long-term QoS for Device-to-Device communications underlaying cellular networks. In [16], the joint beam-vectors design, RRH selection and UE-RRH associations was examined to minimize the total network power consumption for dense C-RAN with incomplete channel state information. However, the

aforementioned works [13]- [16] did not considered the impact of cross-layer interference on the system performance.

Most recently, applying NOMA techniques in the wireless networks has been receiving increasing research attention, due to the fact it can provide high spectral efficiency and boost the user connectivity. In [17], the use of NOMA techniques for short-packet transmissions in the IoT scenarios was examined. In [18], the authors examined the optimal solution that maximizes the weighted sum system throughput of full-duplex NOMA systems. In [19], a novel sub-channel assignment strategy was proposed to maximize the sum rate of small cell users in NOMA-enhanced heterogeneous networks, while taking users' fairness into account. In [20] the joint user association, sub-channel assignment, and power allocation was examined in the multi-cell multi-carrier NOMA systems. However, the performance of NOMA techniques for IoT scenario in the F-RANs without slice isolation has so far remained unclear. Given the significance of NOMA techniques in future wireless networks, it is important to examine the impact of the interactions between different slice instance in the F-RANs with NOMA.

III. System Model and Problem Formulation

Consider the downlink transmission of an F-RAN, as shown in Figure 1, which consists of a BBU pool, S RRHs, indexed by $\mathcal{S}=\{1,2,3,\ldots,S\}$, one FAP, and two kinds of UEs, i.e., the hotspot UEs and the IoT UEs. We assume that the hotspot UEs require high data rate transmissions, while the IoT UEs require transmissions with low end-to-end latency. To this end, the considered F-RAN is divided into two slice instances, namely, the hotspot slice instance and the IoT slice instance. Specifically, the hotspot slice instance is allocated with a BBU pool and S RRHs. Each RRH connects with the BBU pool through the fronthaul links. The RRHs forward the radio signals from the centralized BBU pool to the hotspot UEs (referred to as the RUEs in the sequel). Note that the hotspot slice instance provides high data rate by taking advantage of centralized signal processing and resource allocation in the centralized BBU pool. The IoT slice instance is allocated with one FAP. The FAP is equipped with local storage and processing functionalities, and hence can support low-latency requirements of the IoT UEs (referred to as the FUEs in the sequel).

The FAP and all the UEs are equipped with a single antenna each, while each RRH is equipped with N_s antennas. Let $\mathcal{M}=\{1,2,3,\ldots,M\}$ and $\mathcal{F}=\{1,2,3,\ldots,F\}$ denote the set of all the FUEs and RUEs, respectively. We assume that the F RUEs form T non-overlapping and non-empty multicast groups, indexed by $\mathcal{T}=\{1,2,3,\ldots,T\}$. Let $\mathcal{G}_t$ denote the set of RUEs in the t-th multicast groups. Then, we have $\cup_i \mathcal{G}_i=\mathcal{F}$ and $\mathcal{G}_i \cap \mathcal{G}_j \neq \varnothing$. The total bandwidth of the system, BW, is equally divided into D sub-channels, indexed by $\mathcal{D}=\{1,2,3,\ldots,D\}$. Therefore, the bandwidth of each sub-channel is $\Delta f = BW/D$. In our system, the two slice instances share the same sub-channels.

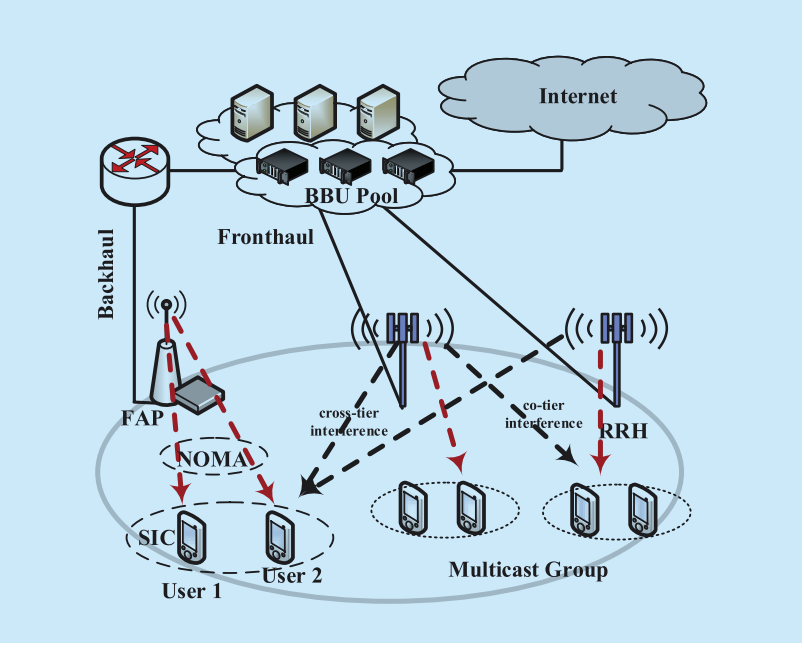


Fig. 1. *Illustration of an F-RAN.*

We denote Γ_m^d as the sub-channel allocation indicator. As such, $\Gamma_m^d = 1$ if sub-channel d is allocated to FUE m and $\Gamma_m^d = 0$ otherwise. We further denote $\mathbf{h}_{fs}^d \in \mathbb{C}^{N_s}, \forall f, s$ and $\mathbf{w}_{st}^d \in \mathbb{C}^{N_s}$ as the propagation channel from the s-th RRH to the f-th RUE on the sub-channel d and the transmit beamforming vector from the s-th RRH to RUEs in the multicast group $\mathcal{G}_t$ on the sub-channel d, respectively.

We first focus on the transmission in the IoT slice instance. In each sub-channel, the FAP serves multiple FUEs simultaneously via the NOMA protocol. We consider a block fading environment, where the channels remain constant within a time slot, but vary independently from one time slot to another. As such, the sub-channel d between the FUE m and the FAP can be denoted by $g_{B,m}^d = \alpha_{B,m} H_{B,m}^d$, where $\alpha_{B,m}$ and $H_{B,m}^d$ represent the path loss coefficient and the fast fading coefficient, respectively. Let $\mathcal{K}_d \in \{\mathcal{K}_1, \mathcal{K}_2, ..., \mathcal{K}_D\}$ and $p_{B,n}^d$ be the set of FUEs allocated on the sub-channel d and the power allocated to the n-th user on the sub-channel d, respectively. Then the sub-channel and FAP power constraints are given by $\sum_{n \in \mathcal{K}_d} p_{B,n}^d = p_B^d$ and $\sum_{d=1}^{D} p_B^d = p_B$, where p_B^d and p_B denote the power allocated on the sub-channel d and the total transmitted power of the FAP, respectively. The received signal at FUE m served by the FAP on sub-channel d is then given by

$$y_{B,m}^d = \underbrace{g_{B,m}^d \sqrt{p_{B,m}^d} u_{B,m}^d}_{\text{desired signal}} + \underbrace{\sum_{n \neq m, n \in \mathcal{K}_d} g_{B,m}^d \sqrt{p_{B,n}^d} u_{B,n}^d}_{\text{interference from NOMA users}} + \underbrace{\sum_{t=1}^{T} \left(\sum_{s=1}^{S} \mathbf{h}_{ms}^{dH} \mathbf{w}_{st}^d \right) u_t^d}_{\text{cross-tier interference}} + n_m^d, \tag{1}$$

where $u_{B,n}^d$ is the symbols transmitted from the FAP to its serving FUE $n \in \mathcal{K}_d$ on the sub-channel d, u_t^d is the transmitted symbol with unit variance for the multicast group t on the sub-channel d in the hotspot slice, n_m^d is the additive white Gaussian noise (AWGN) in the sub-channel d at the UE m with zero mean and variance σ_m^2. Note that the cross-tier interference is due to the transmissions of RRHs to their RUEs on the sub-channel d.

Each FUE m adopts successive interference cancelation (SIC) [21] to decode its desired message from the superimposed signals. Without loss of generality, we assume that the normalized channel gain of FUE $m \in \mathcal{K}_d$ satisfies $|g_{B,1}^d|^2/\sigma_1^2 > |g_{B,2}^d|^2/\sigma_2^2 > \cdots > |g_{B,m}^d|^2/\sigma_m^2$, and choose the decreasing order of the channel gains as the SIC order. As per the rules of NOMA and SIC, the FUE with higher normalized channel gain can cancel the interference from the FUEs with lower normalized channel gains, while the FUE with lower normalized channel gain treats the signals from the FUEs with high normalized channel gains as noise. As such, the signal-to-interference-plus-noise ratio (SINR) of the FUE m over the sub-channel d within one time slot is given by

$$\gamma_{B,m}^d = \frac{p_{B,m}^d |g_{B,m}^d|^2}{I_m^d + A_m^d + \sigma_m^2}, \tag{2}$$

where A_m^d is interference that user m receives from other NOMA FUEs in $\mathcal{K}_d$ in sub-channel d, given by

$$A_m^d = \sum_{n \in \left\{ \mathcal{K}_d \middle| |g_{B,m}^d|^2/\sigma_m^2 > |g_{B,m}^d|^2/\sigma_m^2 \right\}} p_{B,n}^d \left| g_{B,m}^d \right|^2, \tag{3}$$

and I_m^d is the interference that FUE m receives from the transmissions of the hotspot slice on sub-channel d, given by

$$I_m^d = \sum_{t=1}^{T} \left| \sum_{s=1}^{S} \mathbf{h}_{ms}^{dH} \mathbf{w}_{st}^d \right|^2. \tag{4}$$

We now focus on the transmissions in the hotspot slice. It is assumed that all the RRHs have their own transmit power constraints, which form the following feasible set of the beamforming vectors $\mathbf{w}_{st}^d$ on the sub-channel d,

$$\mathcal{W} = \left\{ \mathbf{w}_{st}^d \in \mathbb{C}^{N_s} : \sum_{t=1}^{T} \left\| \mathbf{w}_{st}^d \right\|_2^2 \leq P_{s,max}, \forall s \in \mathcal{S} \right\}, \tag{5}$$

where $P_{s,max} > 0$ is the maximum transmit power of the s-th RRH. We also assume that each RUE is assigned to one sub-channel (as in [6]). Then, the received signal at RUE f in the t-th multicast group on sub-channel d can be written as

$$y_{ft}^{d}=\underbrace{\left(\sum_{s=1}^{S}\mathbf{h}_{fs}^{dH}\mathbf{w}_{st}^{d}\right)u_{t}^{d}}_{\text{desired signal}}+\underbrace{\sum_{i\neq t}\left(\sum_{s=1}^{S}\mathbf{h}_{fs}^{dH}\mathbf{w}_{si}^{d}\right)u_{i}^{d}}_{\text{co-tier interference signal}}+\underbrace{\sum_{m\in\mathcal{K}_{d}}\Gamma_{m}^{d}g_{B,f}^{d}\sqrt{p_{B,m}^{d}}u_{B,m}^{d}}_{\text{cross-tier interference signal}}+n_{f}^{d},\forall f\in\mathcal{G}_{t} \tag{6}$$

Note that the co-tier interference and the cross-tier interference are due to the transmissions from the RRHs to other RUEs and the transmissions from the FAP to its FUEs, respectively. Based on (6), the SINR of RUE f in the t-th multicast group can be expressed as

$$\gamma_{ft}^{d}=\frac{\mathbf{w}_{t}^{dH}\Theta_{f}\mathbf{w}_{t}^{d}}{\sum_{i\neq t}\mathbf{w}_{i}^{dH}\Theta_{f}\mathbf{w}_{i}^{d}+\sum_{m\in\mathcal{K}_{d}}\Gamma_{m}^{d}g_{B,f}^{d}\sqrt{p_{B,m}^{d}}u_{B,m}^{d}+\sigma_{f}^{2}},\quad \forall f\in\mathcal{G}_{t}. \tag{7}$$

In (7), $\Theta_{f}=\mathbf{h}_{f}^{d}\mathbf{h}_{f}^{dH}\in\mathbb{C}^{N\times N}$, where $N=\sum_{s=1}^{S}N_{s}$ and $\mathbf{h}_{f}^{d}=[\mathbf{h}_{f1}^{dH},\mathbf{h}_{f2}^{dH},...,\mathbf{h}_{fs}^{dH}]^{H}$ with $\mathbf{h}_{fs}^{dH}$ denoting the channel vector from the s-th RRH to RUE f, $\mathbf{w}_{t}^{d}=[\mathbf{w}_{1t}^{dH},\mathbf{w}_{2t}^{dH},...,\mathbf{w}_{st}^{dH}]^{H}\in\mathbb{C}^{N}$, where $\mathbf{w}_{st}^{dH}$ denotes the beamforming vector from the s-th RRH to the t-th multicast group. In addition, we define $\mathbf{w}=[\widetilde{\mathbf{w}}_{s}]_{s=1}^{S}\in\mathbb{C}^{TN}$, where $\widetilde{\mathbf{w}}_{s}=[\mathbf{w}_{st}]_{t=1}^{T}\in\mathbb{C}^{TN_{s}}$.

For the hotspot slice instance, the goal is to maximize the number of RUEs that can be supported with desired QoS, while introducing less interference than the maximum interference threshold to the FUEs. Denote $\overline{\mathcal{F}}\subseteq\mathcal{F}$ as the subset of RUEs that are admitted in the hotspot slice, this problem can be formulated as

$$\max_{\{\mathbf{w},\overline{\mathcal{F}}\subseteq\mathcal{F}\}}\quad|\overline{\mathcal{F}}| \tag{8a}$$

$$s.t.\quad \gamma_{ft}^{d}\geq\gamma_{f}^{thr},\forall f\in\overline{\mathcal{F}}, \tag{8b}$$

$$\sum_{t=1}^{T}\left\|\mathbf{w}_{st}^{d}\right\|_{2}^{2}\leq P_{s,\max},\forall s\in\mathcal{S}, \tag{8c}$$

$$\Gamma_{m}^{d}I_{m}^{d}\leq\Gamma_{m}^{d}I_{m,\max}^{d},\forall m\in\mathcal{M}, \tag{8d}$$

where (8b) ensures the QoS of the RUEs, (8c) represents the power budget of the RRHs, and (8d) guarantees that the interference to the FUEs is no more than the maximum interference threshold. In (8d), $I_{m,\max}^{d}$ denotes the maximum tolerable interference threshold at FUE m on sub-channel d. We also note that (8d) is active only if the sub-channel d is allocated to FUE m. For the hotspot slice instance, $I_{m,\max}^{d}$ is a constant in problem (8) that affects the capability of the hotspot slice in supporting the RUEs with desired QoS. However, the value of $I_{m,\max}^{d}$ can be further improved by optimizing the resource allocation strategy of the IoT slice instance.

Note that different resource allocation strategies, employed by the IoT slice instance, will affect the performance of hotspot slice instance. In this paper, different from the traditional "minimize the sum-power" and "maximize the sum-rate" objective functions, we aim to jointly optimize radio resource to maximize the total interference threshold that the FUE can tolerate from the hotspot slice. The motivation behind using this objective function is to provide the largest possibility for the hotspot slice instance in increasing the number of admitted RUEs. The decision process in F-RANs in a resource allocation time slot is as follows. Given the rate requirements for the FUEs, the IoT slice instance first allocates resources to its FUEs and specifies the maximum tolerable interference threshold on each allocated sub-channel. The user pairing and resource allocation results remain fixed throughout the current time slot. The FAP then sends those resource allocation results to the BBU pool. The BBU pool then centrally performs resource allocation and admission control for their RUEs in the hotspot slice instance. In the end, the FUEs perform interference measurements to verify the feasibility of the scheme. According to (8d), we note that the performance of the hotspot slice instance is constrained by the interference threshold that the FUEs can tolerate. In order to maximize the capability of the hotspot slice in supporting the RUEs with desired QoS, we further jointly optimize the sub-channel and power allocation to maximize the total interference threshold that the FUE can tolerate from the hotspot slice instance, while satisfying the QoS constraints of the FUEs. This optimization problem in the IoT slice instance can be formulated as

$$\max_{\{\Gamma_m^d, I_{m,\max}^d, p_{B,m}^d\}} \quad \sum_{m=1}^{M}\sum_{d=1}^{D} I_{m,\max}^d \tag{9a}$$

$$s.t. \quad \sum_{d=1}^{D} \Delta f \log_2\left(1+\frac{p_{B,m}^d \mid g_{B,m}^d \mid^2}{I_{m,\max}^d + A_m^d + \sigma_m^2}\right) \geq R_m, \tag{9b}$$

$$\forall m \in \mathcal{M},$$

$$\sum_{m=1}^{M} p_{B,m}^d \leq p_{B,\max}^d, \forall d \in \mathcal{D}, \tag{9c}$$

$$\sum_{m=1}^{M} \Gamma_m^d \leq 2, \forall d \in \mathcal{D}, \tag{9d}$$

$$\sum_{d=1}^{D} \Gamma_m^d \leq 1, \forall m \in \mathcal{M}, \tag{9e}$$

$$p_{B,m}^d \geq 0, \forall m \in \mathcal{M}, d \in \mathcal{D}, \tag{9f}$$

$$I_{m,\max}^d \geq I_{\min}, I_{m,\max}^d \leq I_{\max}, \forall m \in \mathcal{M}, d \in \mathcal{D} \tag{9g}$$

where the objective is to maximize the sum of the tolerable interference threshold for all the FUE on all the sub-channels, (9b) denotes minimum data rate requirement for each FUE, (9c) is the power constraint for the FAP with maximum transmit power allowance $p_{B,\max}^d$ on sub-channel d. To reduce hardware complexity and processing delay, (9d) and (9e) ensure that each sub-channel can only be assigned to at most two users via the NOMA protocol, and each user can only occupy at most one sub-channel. Note that user pairing is performed on each subcarrier. (9f) is the non-negative transmit power constraint. Finally, (9g) sets lower and upper bounds on the value of $I_{m,\max}^d$. Note that the value of $I_{\max}$ is relatively large. As such, $I_{m,\max}^d = I_{\max}$ indicates that the sub-channel d is not assigned to an FUE m since there is no restriction on the threshold of the interference in the sub-channel d.

IV. Resource Allocation and User Pairing Algorithm for IOT Slice Instance

In this section, we propose a low-complexity algorithm to deal with Problem (9) for the IoT slice instance. Note that the problem in (9) is a mixed integer nonlinear programming problem, which is mathematically intractable. To proceed, we decouple the problem in (9) to a sub-channel allocation problem and a power allocation problem. Specifically, for a fixed power allocation, the sub-channel allocation problem can be formulated as a many-to-one two-sided matching problem. For a given sub-channel allocation, the power allocation problem can be formulated as a linear programming problem. As such, the sub-channel allocation problem and power allocation problem can be solved by using the matching games and the interior point methods, respectively.

4.1 Sub-channel allocation

In this subsection, we focus on resolving the sub-channel allocation problem. To this end, we first assume that (9) is feasible (as in [22]) and that an FUE can have its rate requirement satisfied with one sub-channel only. This assumption is practical in scenarios where the FAP can control the allowable interference threshold $I_{m,\max}^d$ on the sub-channel d allocated to FUE m. Based on this assumption, we obtain a feature of problem (9), as follows:

Lemma 1: At optimality, the data rate constraints for the FUEs (i.e., 9b) holds with equality.

Proof: Since the objective function in (9) monotonically increases as $I_{m,\max}^d$ whereas the constraint (9b) monotonically decreases as $I_{m,\max}^d$, (9b) must hold with equality at optimality for the FUE.

Based on Lemma 1, the utility of the FUE m, which is defined as the tolerable interference threshold of FUE m, is expressed as

$$U_m = \frac{g_{B,m}^d p_{B,m}^d}{2^{R_m/\Delta f}-1} - A_m^d - \sigma_m^2, \tag{10}$$

and the utility of sub-channel d is given by $U_d = \sum_{m \in \mathcal{K}_d} U_m$.

Then, we consider the sub-channel allocation as a two-sided matching process Ψ between the set of the FUEs (i.e., $\mathcal{M} = \{1,2,3,\ldots,M\}$) and the set of the sub-channels (i.e., $\mathcal{D} = \{1,2,3,\ldots,D\}$). As such, Ψ should satisfy the following definition.

Definition 1: For two disjoint sets, $\mathcal{M} = \{1,2,3,\ldots,M\}$ and $\mathcal{D} = \{1,2,3,\ldots,D\}$, a many-to-one matching Ψ is a mapping from the set $\mathcal{M} \cup \mathcal{D}$ into the set of all sub-

sets of $\mathcal{M} \cup \mathcal{D}$ such that 1) $\Psi(m) \subseteq \mathcal{D}$; 2) $\Psi(d) \subseteq \mathcal{M}$; 3) $|\Psi(d)| \leq 2$; 4) $|\Psi(m)| = 1$; 5) $\Psi(m) = d \Leftrightarrow m \in \Psi(d)$.

In Definition (1), conditions 1) and 2) imply that each FUE is matched with a subset of sub-channels and each sub-channel is matched with a subset of FUEs. Taking into account the complexity of the decoding technique at the receiver, we set the sizes of $\Psi(d)$ and $\Psi(m)$ to be no larger than 2 and 1, respectively (as expressed in conditions 3) and 4)). In addition, condition 5) implies that if the FUE m is matched with the sub-channel d, then the sub-channel d is also matched with FUE m.

To start the matching process, both the FUEs and sub-channels need to set up the preference lists according to their own interests. Specifically, compared to sub-channel d', if FUE m has a higher utility when associated with sub-channel d, i.e., $U_m(d) > U_m(d')$, we have $d \succ_m d'$, indicating that FUE m prefers d to d'. Each FUE m forms a preference list $\mathcal{FP}_d$ with the descending order of the utility. Since each sub-channel can be matched with up to 2 FUEs, if $U_d(\mathcal{K}_d) > U_d(\mathcal{K}'_d) \Rightarrow \mathcal{K}_d \succ_d \mathcal{K}'_d$, the sub-channel d prefers the set of FUEs $\mathcal{K}_d$ to $\mathcal{K}'_d$. Similarly, each sub-channel d forms a preference list $\mathcal{SP}_m$ over all the possible sets of FUEs with the descending order of the utility.

After the FUEs and sub-channels construct their own preference lists, the matching process can be performed via the *swap operation* [23]. To do so, we define

$$\begin{aligned}\Psi_m^{m'} = &\{\Psi \setminus \{(m, \Psi(m)), (m', \Psi(m'))\}\} \\ &\cup \{(m, \Psi(m')), (m', \Psi(m))\},\end{aligned} \tag{11}$$

and present the definition of swap-blocking pair, as follows:

Definition 2: A pair of FUEs (m, m') is a swap-blocking pair if and only if

1) $\forall r \in \{m, m', \Psi(m), \Psi(m')\}, U_r(\Psi_m^{m'}) \geqslant U_r(\Psi)$ and

2) $\exists r \in \{m, m', \Psi(m), \Psi(m')\}$, such that $U_r(\Psi_m^{m'}) > U_r(\Psi)$, where $U_r(\Psi)$ denotes the utility of the player r under the matching state Ψ.

In Definition 2, condition 1) implies that the tolerable interference threshold of any players involved will not be reduced after the swap operation and condition 2) indicates that at least one of the players' tolerable interference threshold will increase after the swap operation.

1) Proposed Sub-channel Allocation Algorithm: Based on the above analysis, in this subsection, a suboptimal FUE-sub-channel matching algorithm (FSMA) is proposed, as shown in Algorithm 1. The key idea of FSMA is to keep considering approved swap matchings among the players so as to reach a two-sided exchange stable matching. In the initialization phase, a priority-based allocation scheme is applied **(Step 2)**. Each FUE chooses its preferred set of available sub-channels **(Step 5)**. This process continues until the set $\mathcal{K}_u$ goes empty **(Step 8)**. The swap matching phase contains multiple iterations in which the FAP keeps searching for two FUEs to form a swap-blocking pair and update the current matching **(Step 10-Step 18)**. The iterations stop until there exists no swap-blocking pair and a final matching is determined **(Step 19)**. In addition, to prevent meaningless swap operations between FUE m and m', we define a variable $\mathcal{SR}_{m,m'}$ to record the times that the FUE swaps their former matched sub-channels. Due to the use of swap operations, the convergence of FSMA can be confirmed [24].

2) Complexity Analysis: In the FSMA, the computational complexity mainly comes from two phase, namely, the initialization phase and the swap matching phase. For the initialization phase, the complexity of setting up the preference lists of the FUEs and sub-channels is $O(MD^2)$. In each iteration of swap matching phase, at most $M(D-1)$ swap matchings need to be considered when $M = 2D$. Given the number of iterations ς, the computational complexity of FSMA can be approximated by $O(\varsigma 2MD)$. Note that, although the swap matching is guaranteed to converge, the total number of iterations needed for the convergence cannot be analytically obtained. In Section VI, we will numerically examine the impact of the number of the FUEs on the number

of iterations in the swap matching phase.

4.2 Power allocation

In this subsection, we focus on the power allocation at the FAP for a fixed sub-channel allocation, i.e.,Γ_m^d. To this end, we re-express the problem in (9) as

$$\max_{\{I_{m,\max}^d, p_{B,m}^d\}} \sum_{m=1}^{M} \sum_{d\in\mathcal{D}_m} I_{m,\max}^d$$
$$s.t.\ C1: p_{B,m}^d \mid g_{B,m}^d \mid^2 \geq$$
$$(I_{m,\max}^d + A_m^d + \sigma_m^2)(2^{R_m/\Delta f} - 1), \forall m \in \mathcal{M}, d \in \mathcal{D}_m,$$
$$C2: \sum_{m=1}^{M} p_{B,m}^d \leq p_{B,\max}^d, \forall d \in \mathcal{D}_m \quad (12)$$
$$C3: p_{B,m}^d \geq 0, \forall m \in \mathcal{M}, d \in \mathcal{D}_m$$
$$C4: I_{m,\max}^d \geq I_{\min}, \forall m \in \mathcal{M}, d \in \mathcal{D}_m$$

In (12), we set $I_{m,\max}^d = I_{\max}$ and $p_{B,m}^d = 0$ for the unassigned sub-channels. Recall that the problem (12) is a linear program, which can be optimally solved by using the interior-point method [25].

Algorithm 1. FUE-sub-channel matching algorithm (FSMA).

1: ***Initialization Phase***
2: Construct the preference lists of the FUEs and the sub-channels, i.e., $\mathcal{FP}_m$ and $\mathcal{SP}_d$; Construct the set of the FUEs that are not matched $\mathcal{K}_u$;
3: **while** $\mathcal{K}_u \neq \varnothing$ **and** $\exists \mathcal{FP}_m \neq \varnothing$ **do**
4: **for** $\forall m \in \mathcal{K}_u$ **do**
5: FUE m matches with its most preferred sub-channel which is not fully matched;
6: Remove m from $\mathcal{K}_u$;
7: **end for**
8: **end while**
9: ***Swap matching phase***
10: Initialize the number of swapping requests that FUE m sends to m', i.e.,$\mathcal{SR}_{m,m'} = 0$;
11: For each FUE m, it searches for another FUE m' to check whether it is a swap-blocking pair;
12: **if** (m, m') forms a swap-blocking pair along with $d = \Psi(m), d' = \Psi(m')$, and $\mathcal{SR}_{m,m'} + \mathcal{SR}_{m',m} < 2$, **then**
13: Update the current matching state to $\Psi_m^{m'}$;
14: $\mathcal{SR}_{m,m'} = \mathcal{SR}_{m,m'} + 1$;
15: **else**
16: Keep the current matching state;
17: **end if**
18: **Repeat** *Step 9*-*Step 17* until there is no swap-blocking pair.
19: ***End of the algorithm***

Algorithm 2. Joint sub-channel and power allocations algorithm (JSPA).

1: Step 1: ***Initialization Phase***
2: The FAP allocates the transmitted power equally to each user over each sub-channel.
3: Set i=0.
4: **Step 2:** ***Joint Sub-channel and Power Allocation***
5: **Repeat**
6: Update the sub-channel allocation result Γ_m^d by using Algorithm 1.
7: Update $p_{B,m}^d$ by solving linear program formulated in (12).
8: Set i=i+1.
9: **Until** convergence.
10: **Step 3:** ***End of the algorithm***

4.3 Joint sub-channel and power allocation

Based on the analysis in Sections IV-A and IV-B, the joint resource allocation algorithm is proposed for the problem in (9), as shown in Algorithm 2. Specifically, Algorithm 2 consists of the initialization phase and the resource allocation phase. In the initialization phase, the FAP allocates the transmitted power equally to each user over each sub-channel. In the resource allocation phase, the sub-channel and power allocation are iteratively performed so as to obtain a joint solution. Note that, due the the upper bound on the total tolerable interference threshold for the FUEs, Algorithm 2 is guaranteed to converge. In the following, we will analyze the complexity of Algorithm 2. In each iteration, the FSMA algorithm is adopted in step 6 with complexity given by $O(\varsigma 2MD)$ as shown in the above subsection, and interior-point method is adopted in step 7 with complexity given by $O((2M)^{3.5})$. Denote ς_J as the average iteration number, then the computational complexity of Algorithm 2 is $O(\varsigma_J \varsigma 2MD + \varsigma_J (2M)^{3.5})$.

V. Admission Control and Resource Allocation Algorithm for Hotspot Slice Instance

After obtaining the maximum total interference threshold for the FUEs, in this section,

the number of RUEs with desired QoS satisfied is maximized, while introducing the interference to the FUEs no more than the maximum total interference threshold. We consider that each sub-channel is pre-allocated to one RUE. However, the hotspot slice instance may not be able to support all the RUEs with their desired QoS, due to the constraints (8c) and (8d). In order to make the optimization problem in (8) feasible, we adopt the admission control to temporarily drop some RUEs. In the following, we first transform the optimization problem into an equivalent form and then solve it by using a two-stage optimization method.

5.1 Problem reformulation

As per the rules of admission control, we introduce an F-dimensional nonnegative real vector $\mathbf{x}=[x_1,x_2,\cdots,x_F]$ into the constraint (8b) and obtain

$$\gamma_{ft}^{d}-\gamma_{f}^{thr}\geq x_f,\forall f\in\mathcal{G}_t. \tag{13}$$

Note that (13) is equivalent to (8a) only when $x_f=0$. That is, $x_f=0$ indicates that the f-th RUE can be admitted. Substituting (7) into (13), we can rewrite (13) as the quadratic constraint, expressed as

$$\begin{aligned}\Phi_{f,t}^{d}(\mathbf{w})=\gamma_{f}^{thr}\Bigg(&\sum_{i\neq t}\mathbf{w}_i^{dH}\Theta_f\mathbf{w}_i^{d}+\\&\Gamma_m^d g_{B,f}^{d}\sqrt{p_{B,m}^{d}}u_{B,m}^{d}+\sigma_f^2\Bigg)\\&-\mathbf{w}_t^{dH}\Theta_f\mathbf{w}_t^{d}\leq x_f,\forall f\in\mathcal{G}_t\end{aligned} \tag{14}$$

As such, maximizing the number of admitted RUEs is equivalent to minimize the number of non-zero x_f's. The optimization problem in (8) is reformulated as

$$\min_{\{\mathbf{x},\mathbf{w}\}}\quad \|\mathbf{x}\|_0 \tag{15a}$$

$$s.t.\quad \Phi_{f,t}^{d}(\mathbf{w})\leq x_f,\forall f\in\mathcal{G}_t, \tag{15b}$$

$$\sum_{t=1}^{T}\left\|\mathbf{w}_{st}^{d}\right\|_2^2\leq P_{s,\max},\forall s\in\mathcal{S}, \tag{15c}$$

$$\Gamma_m^d I_m^d\leq\Gamma_m^d I_{m,\max}^d,\forall m\in\mathcal{M}. \tag{15d}$$

The optimization problem in (15) is still mathematically intractable due to the non-convex objective function and the quadratic constraint. To address this issue, a two-stage optimization method is adopted. Specifically, we first convexify the quadratic QoS constraint in (15b) by using the semi-definite relaxation technique (SDR) [26]. Then, l_p-norm minimization approach is used to approximate the l_0 -norm based problem.

5.2 Algorithm design

We first re-express the non-convex QoS constraint in (14) as

$$\begin{aligned}\Phi_{f,t}^{d}(\mathbf{Q})=\gamma_{f}^{thr}\Bigg(&\sum_{i\neq t}\mathrm{Tr}(\Theta_f\mathbf{Q}_i)+\\&\Gamma_m^d g_{B,f}^{d}\sqrt{p_{B,m}^{d}}u_{B,m}^{d}+\sigma_f^2\Bigg)\\&-\mathrm{Tr}(\Theta_f\mathbf{Q}_t)\leq x_f,\end{aligned} \tag{16}$$

where $\mathbf{Q}=[\mathbf{Q}_t]_{t=1}^{T}$ with $\mathbf{Q}_t^d=\mathbf{w}_t^d\mathbf{w}_t^{dH}\in\mathbb{C}^{N\times N}$ and the rank of $\mathbf{Q}_t$ is one.

Based on (16), the total power constraint for each RRH in (15c) can be rewritten as

$$\mathcal{W}=\left\{\mathbf{Q}_t^d\in\mathbb{C}^N:\sum_{t=1}^{T}\mathrm{Tr}(\mathbf{C}_{st}\mathbf{Q}_t^d)\leq P_{s,\max},\forall s\in\mathcal{S}\right\}, \tag{17}$$

where $\mathbf{C}_{st}\in\mathbb{R}^{N\times N}$ is a block diagonal matrix, where the s-th main diagonal block square matrix is the identity matrix $\mathbf{I}_{N_s}$ and other elements are zeros. Note that the constraints in (16) and (17) are convex.

We then focus on convexifying the objective function in (15a). Specifically, we approximate the non-convex l_0-norm in (15a) by a smoothed version of l_p-norm $\|\mathbf{x}\|_p^p$ to seek sparser solutions [27], i.e.,

$$f_p(\mathbf{x};\epsilon):=\sum_{f=1}^{F}(x_f^2+\epsilon^2)^{p/2}, \tag{18}$$

where $\mathbf{x}\in\mathbb{R}^F$ and $\epsilon>0$ is a small fixed regularizing parameter. We note that such approximation does not change the optimal solution when ϵ is small.

Based on (16)–(18), the optimization problem in (15) can be further transformed as

$$\min_{\{\mathbf{x},\mathbf{Q}\}}\quad \sum_{f=1}^{F}(x_f^2+\epsilon^2)^{p/2} \tag{19a}$$

$$s.t.\quad \Phi_{f,t}^{d}(\mathbf{Q})\leq x_f,\forall f\in\mathcal{G}_t, \tag{19b}$$

$$\sum_{t=1}^{T}\mathrm{Tr}(\mathbf{C}_{st}\mathbf{Q}_t^d)\leq P_{s,\max},\forall s\in\mathcal{S}, \tag{19c}$$

$$\Gamma_m^d\left(\sum_{t=1}^{T}\mathrm{Tr}(\Theta_m^d\mathbf{Q}_t^d)\right)\leq\Gamma_m^d I_{m,\max}^d,\forall m\in\mathcal{K}, \tag{19d}$$

$$\mathbf{Q}_t\succ 0,\forall t\in\mathcal{T}, \tag{19e}$$

Note that, as per the rules of the SDR technique, the rank-one constraints for all the $\mathbf{Q}_t^d$ are dropped in (19). As such, the constraints in (19) are all convex. The iteratively reweighted least squares (IRLS) approach [28] can be used to solve (19), as follows. Given an initial feasible solution $\mathbf{x}^{[0]}$ of problem, the IRLS approach generates a sequence $\{\mathbf{x}^{[n]}\}_{n=1}^{\infty}$ by successively minimizing the upper bounds $L(\mathbf{x},\boldsymbol{\omega}^{[n]}) = \sum_{f=1}^{F} \omega_f^{[n]} x_f^2$ of the objective function $f_p(\mathbf{x};\epsilon)$. Let $\mathbf{x}^{[n+1]}$ be the minimizer of the upper bound function at the n-th iteration. Mathematically, it is expressed as

$$\mathbf{x}^{[n+1]} := \arg\min \sum_{f=1}^{F} \omega_f^{[n]} x_f^2, \tag{20}$$

where

$$\omega_f^{[n]} = \frac{p}{2}\left[\left(x_f^{[n]}\right)^2 + \epsilon^2\right]^{\frac{p}{2}-1}, \forall f = 1,...,F. \tag{21}$$

Based on the IRLS approach, the problem in (19) is solved using Algorithm 3. In the following proposition, we prove the convergence of Algorithm 3.

Proposition 1: Algorithm 3 guarantees to converge.

Proof: In order to prove the convergence of the Algorithm 3, we first define the approximation error function as

$$\begin{aligned} G(\mathbf{x}) &= f_p(\mathbf{x};\epsilon) - L(\mathbf{x},\boldsymbol{\omega}^{[n]}) \\ &= \sum_{f=1}^{F} (x_f^2 + \epsilon^2)^{p/2} - \sum_{f=1}^{F} \omega_f^{[n]} x_f^2, \end{aligned} \tag{22}$$

which is concave in $\mathbf{x} \in \mathbb{R}^F$. By solving the equation $\nabla_{\mathbf{x}} G(\mathbf{x}) = 0$, we can easily obtain that the function $G(\mathbf{x})$ attains the maximum at $\mathbf{x} = \mathbf{x}^{[n+1]}$. Based on above equation and the definition of $\mathbf{x}^{[n+1]}$ in (20), we have

$$\begin{aligned} &f_p(\mathbf{x}^{[n+1]};\epsilon) - f_p(\mathbf{x}^{[n]};\epsilon) = f_p(\mathbf{x}^{[n+1]};\epsilon) \\ &\quad - f_p(\mathbf{x}^{[n]};\epsilon) + L(\mathbf{x}^{[n+1]},\boldsymbol{\omega}^{[n]}) - L(\mathbf{x}^{[n+1]},\boldsymbol{\omega}^{[n]}) \\ &\le f_p(\mathbf{x}^{[n]};\epsilon) - L(\mathbf{x}^{[n]},\boldsymbol{\omega}^{[n]}) \\ &\quad - f_p(\mathbf{x}^{[n]};\epsilon) + L(\mathbf{x}^{[n+1]},\boldsymbol{\omega}^{[n]}) \\ &= L(\mathbf{x}^{[n+1]},\boldsymbol{\omega}^{[n]}) - L(\mathbf{x}^{[n]},\boldsymbol{\omega}^{[n]}) \le 0 \end{aligned} \tag{23}$$

where the first inequality is based on the fact that function $G(\mathbf{x})$ attains the maximum at $\mathbf{x} = \mathbf{x}^{[n+1]}$, and the second inequality is based on the fact that $\mathbf{x}^{[n+1]}$ be the minimizer of the upper bound function at the n-th iteration. Thus, the value of the objective function in the *(n+1)*th iteration is less than or equal to that in the *n*th iteration. That is, Algorithm 3 produces a sequence of non-increasing values of the objective function. In addition, the value of the objective function is bounded by the constraints in (19). As such, we note that Algorithm 3 converges to the local optimal solution of (19). Futhermore, based on the analysis in [14] and [28], as well as expectation-maximization theory, the solution of the Algorithm 3 satisfies the Karush-Kuhn-Tucker (KKT) conditions of the original problem (19), which verifies the convergence of Algorithm 3. The proof is completed.

Denote the solution of Algorithm 3 as $\mathbf{x}^*$. If $x_f = 0, \forall f \in \mathcal{G}_t$, all the RUEs can be admitted in the network. Otherwise, some RUEs should be removed. Intuitively, the RUE with a smaller x_f should have a higher priority to be admitted since it has the smaller gap between the target SINR and the achievable SINR for RUE f. We sort the coefficients in the descending order: $x_{\pi_1} \ge x_{\pi_2} \ge \cdots \ge x_{\pi_F}$, where π_f denotes the RUE that ranks fth in the descending order. Then, admitting the maximum number of RUEs is equivalent to find a minimum J_0 such that all the remaining RUEs in $\mathcal{U} = \{\pi_{J_0+1}, \cdots, \pi_F\}$ can be supported. The bisection search approach can be used to find the minimum J_0. Specifically, we check whether all the RUEs in $\mathcal{U}$ can be supported in each iteration. That is, we need to solve the following feasibility problem

$$\text{find} \{\mathbf{Q}_t^d\}_{t \in \mathcal{T}} \tag{24a}$$

$$s.t. \quad \Phi_{f,t}^d(\mathbf{Q}) \le 0, \forall f \in \mathcal{U}, \tag{24b}$$

$$(19c) - (19e) \tag{24c}$$

If problem (24) is feasible, it implies that

Algorithm 3. The approach to solve the problem in (19).

1: **Initialize** Find an initial feasible point $\mathbf{x}^{[0]}$ and set $\boldsymbol{\omega}^{[0]} = (1,...,1)$ and k=0.
2: **repeat**
3: 1) Compute the optimal solution $\mathbf{x}^{[n+1]}$ according to (20).
4: 2) Update the weights according to (21).
5: Set $n=n+1$.
6: **until** Convergence.

the QoS constraints of all the RUEs in $\mathcal{U}$ can be satisfied. We denote the corresponding sets of admitted UEs and the set of multicast groups as $\mathcal{F}^*$ and $\mathcal{T}^*$, respectively. We have $\mathcal{T}^* = \{t : \mathcal{G}_t \cap \mathcal{F}^* \neq \varnothing\}$. Then, the following transmission power minimization problem needs to be solved as well

$$\min_{\{\mathbf{Q}\}} \quad \sum_{s=1}^{S} \sum_{t \in \mathcal{T}^*} \mathrm{Tr}(\mathbf{C}_{st} \mathbf{Q}_t^d) \tag{25a}$$

$$s.t. \quad \Phi_{f,t}^d(\mathbf{Q}) \leq 0, \forall f \in \mathcal{F}^*, \tag{25b}$$

$$(19c) - (19e) \tag{25c}$$

Note that the problem (25) is a semi-definite programming (SDP) problem and can be solved by using the standard convex optimization methods such as classic interior point method or CVX [29]. Recall that we drop the rank one constraints for all the $\mathbf{Q}_t$, therefore the solution of (25) should be checked if the rank one constraints are satisfied. Specifically, the Gaussian randomization method can be employed to obtain the feasible rank-one approximate solution [30] if the constraints are not satisfied. If the $\mathbf{Q}_t$ is not rank-one, then the BBU pool applies the eigen-value decompsotion on $\mathbf{Q}_t$ to get $\mathbf{Q}_t = \mathbf{U}_t \Sigma_t \mathbf{U}_t^H$, where $\mathbf{U}_t = [e_1, ..., e_N]$ and $\Sigma_t = diag(\lambda_1, ..., \lambda_N)$ a unitary matrix and a diagonal matrix, respectively. Then, we obtain a suboptimal solution as $\mathbf{w}_t = \mathbf{U}_t \Sigma_t^{1/2} \mathbf{r}$, where $\mathbf{r}$ is a random vector generated according to the circularly symmetric complex Gaussian distribution with zero mean and unit variance. In Algorithm 4, we summarize the joint admission control and power minimization algorithm that solves (15).

5.3 Complexity analysis

In this subsection, we examine the computational complexity of Algorithm 4. Note that the optimal admission control can only be obtained by the exhaustive search over all the possible subsets of the RUEs, the computational complexity of which increases exponentially with the number of the RUEs. In our Algorithm 4, the feasibility problem in (24) can be solved within no more than $(1+\log(1+F))$ iterations. In addition, the power minimization SDP problem in (25) has T matrices with the size of $N \times N$ each and $(F+S+2)$ linear constraints. Therefore, the corresponding computational complexity is $O(T^3 N^6 + (F+S+2)TN^2)$ per iteration, and the interior-point method usually converges within a few tens of iterations.

VI. Numerical Results and Discussions

In this section, numerical results are presented to validate the effectiveness of our proposed algorithms. Multiple FUEs and RUEs are considered to be randomly distributed in the 0.2km×0.2km hotspot area. For the downlink NOMA network, the hardware complexity and processing delay increases with the number of users multiplexed on the same subchannel. Furthermore, in practice, SIC decoders are imperfect and the resulting decoding errors propagate over multiple decoding stages, leading to additional performance degradation. For simplicity, it is assumed that each sub-channel is only allocated to two FUEs via the NOMA protocol. However, our proposed algorithms can be readily extended to the cases of multiple FUEs sharing the same sub-channel. The minimum rate requirements for the FUEs of poor and good channel conditions are denoted by R_m^{BU} and R_m^{GU}, respectively. In addition, The channel path-loss is modeled as $38.5 + 40.0 * log_{10}(R)$, where R denotes the

Algorithm 4. Joint admission control and power minimization algorithm.

1: Solve the problem (19) by Algorithm 3. Obtain the solution $\mathbf{x}^*$ and sort the entries in the descending order: $x_{\pi_1} \geq ... \geq x_{\pi_F}$.
2: Initialize $J_{\min} = 0, J_{\max} = F, i = 0$.
3: **repeat**
4: 1) Set $i \leftarrow \left\lfloor \frac{J_{min} + J_{max}}{2} \right\rfloor$
5: 2) Solve problem (24a) with $\mathcal{U} = \{\pi_{i+1}, ..., \pi_F\}$: if it is feasible, set $J_{\max} = i$; otherwise, set $J_{\min} = i$.
6: **until** $J_{\max} - J_{\min} = 1$, obtain $J_0 = J_{\max}$ and obtain the admitted RUE set $\mathcal{U}^* = \{\pi_{J_0+1}, ..., \pi_F\}$.
7: Solve problem (25) using the interior point methods to obtain the minimum total transmission power.
8: **end**

Table I. *System parameters.*

Maximum transmit power of FAP $p_{B,\max}$	46 dBm
Maximum transmit power of RRH $p_{s,\max}$	30 dBm
Fixed regularizing parameter ϵ	10^{-3}
Noise power spectral density σ^2	-174 dBm/Hz
The system bandwidth BW	1.8MHz
Noise power of a single subchannel σ_m^2	$\frac{BW}{D}\sigma^2$
Minimal tolerable interference threshold $I_{\min}$	σ_m^2
Maximal tolerable interference threshold $I_{\max}$	$10^5\sigma_m^2$

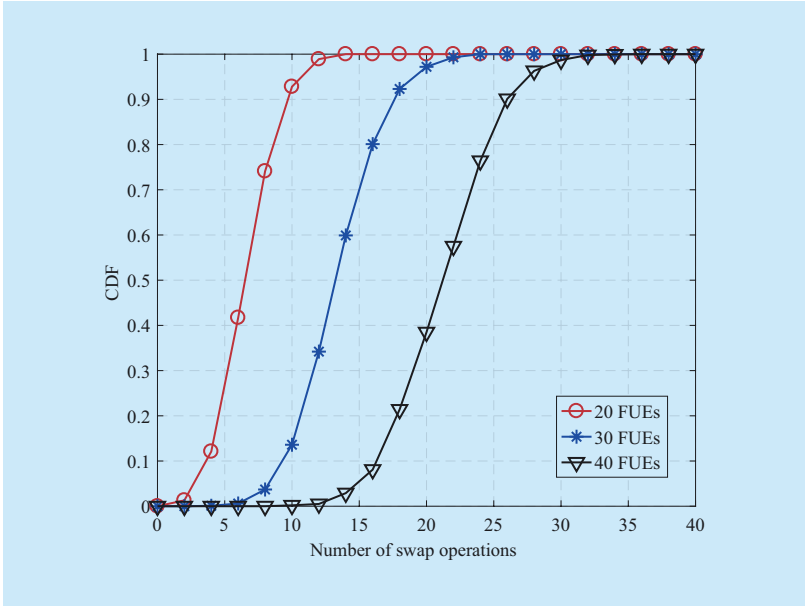


Fig. 2. *CDF of the number of swap operations in FSMA.*

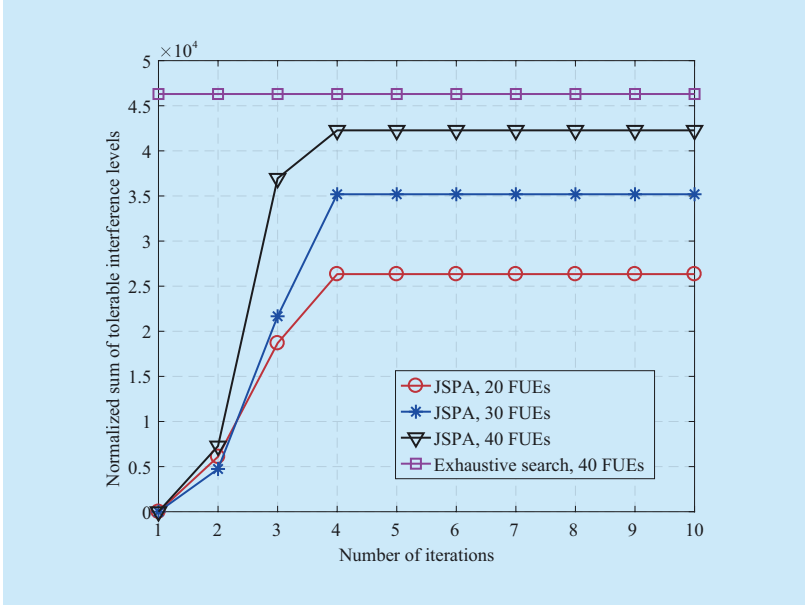


Fig. 3. *Convergence of JSPA for different number of FUEs with* $R_m^{BU} = 2$ *bps/Hz and* $R_m^{GU} = 4$ *bps/Hz.*

distance between the access points and the UE (in meters) [31]. The small-scale fading coefficients are modeled as independent complex Gaussian random variables with zero mean and unit variance. Other parameters used in the simulations are summarized in Table I, unless otherwise specified.

6.1 Performance of IoT slice instance

1) Convergence of Proposed FSMA and JSPA for IoT slice instance: In Figure 2 and Figure 3, the convergence of the proposed FSMA and JSPA is verified. Specifically, Figure 2 depicts the cumulative distribution function (CDF) of the number of swap operations for the matching process for different number of FUEs. It is shown that the proposed FSMA can converge within 35 iterations when the number of FUEs ranges from 20 to 40, thus verifying the low computational complexity of the proposed FSMA. In addition, it is observed that the number of swap operations increases as the number of FUEs increases. This is because the probability of there exists the swapping pairs increases when the number of FUEs becomes relatively large.

Figure 3 depicts the sum of tolerable interference threshold (on the allocated sub-channels only, i.e., those sub-channels with non-zero power) versus the number of iterations in the proposed JSPA for different number of FUEs with $R_m^{BU} = 2$bps/Hz and $R_m^{GU} = 4$bps/Hz. For comparison purpose, we provide the optimal performance for the case of 40 FUEs, which is obtained from the exhaustive search in the space of all the possible sub-channel assignment and power allocation. It is seen that the proposed JSPA can converge within a few iterations for different number of FUEs. It is also seen that, when the number of FUEs is 40, the performance of the proposed JSPA is more than 90% of the optimal performance. Moreover, it is observed that a larger sum of tolerable interference threshold can be achieved with a larger number of FUEs.

2) Impact of System Parameters: In Figure

4, the impact of system parameters (i.e., R_m^{GU} and p_B) on the performance of the IoT slice instance is examined. Specifically, we plot the normalized sum of tolerable interference threshold versus the data rate requirement of FUEs with a better channel condition, R_m^{GU}, for different total transmit power at the FAP, p_B. In this figure, it is considered that $R_m^{BU} = 1$bps/Hz and the number of FUEs is $M = 20$. It can be seen that the sum of tolerable interference threshold decreases as R_m^{GU} increases. In addition, it is observed that the sum of tolerable interference threshold increases significantly as p_B.

6.2 Performance of hotspot slice instance

In Figure 5 and Figure 6, the performances of different UE selection algorithms, i.e., the proposed IRLS approach, the MDR approach,

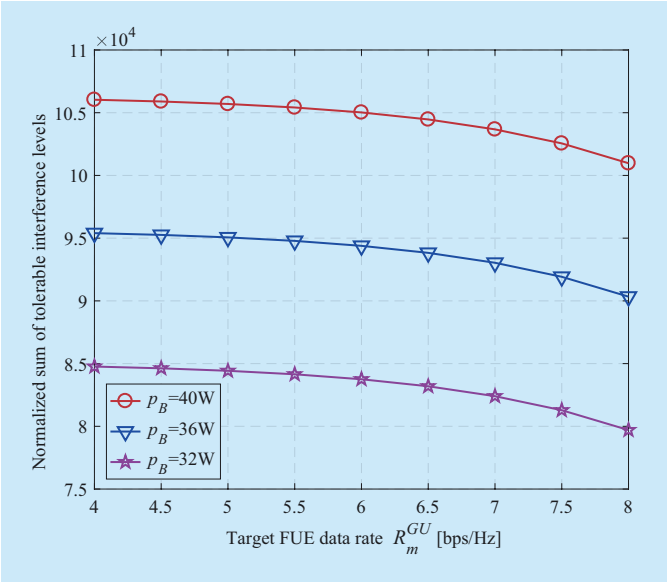


Fig. 4. *Normalized sum of tolerable interference threshold versus* R_m^{GU}.

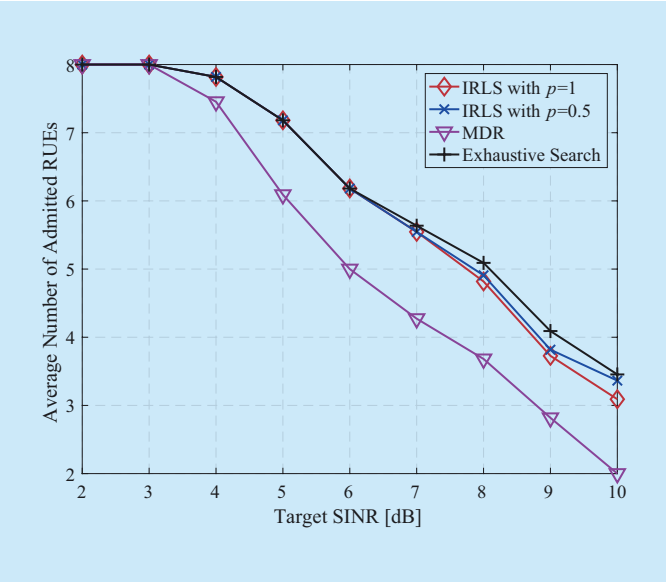


Fig. 5. *The average number of admitted RUEs versus target SINR of RUEs for different algorithms.*

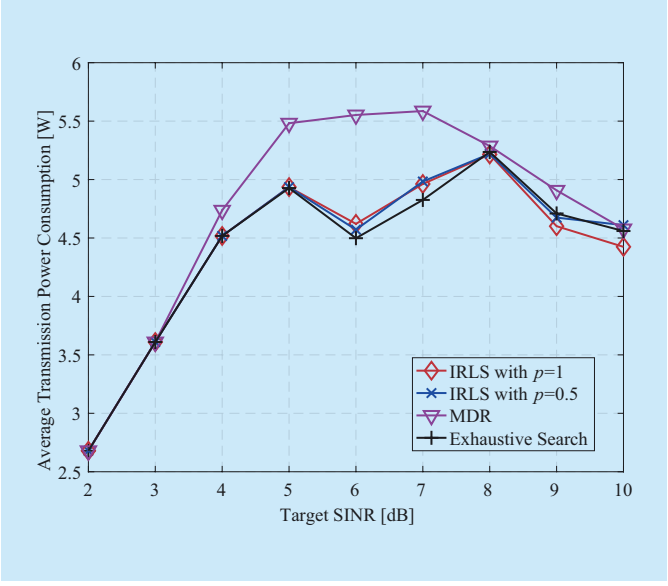


Fig. 6. *The average transmission power versus target SINR of RUEs for different algorithm.*

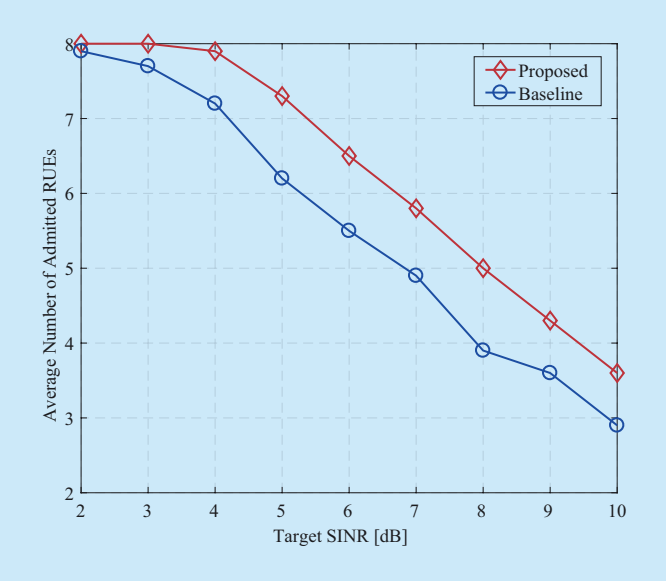


Fig. 7. *The average number of admitted RUEs versus target SINR for different FAP resource allocation algorithms.*

and the exhaustive search, are examined. In particular, the MDR approach refers to the multicast membership deflation by the relaxation algorithm in [32]. In these figures, there exist 6 two-antenna RRHs and 4 multicast groups with 2 single-antenna RUEs in each group, i.e., $S = 6$ and $T = 4$. In addition, we consider a probabilistic channel model in which the large scale channel gain from RUE f to RRH s satisfies: $\alpha_{fs} = 1, \forall s \in \Omega_1$ with $|\Omega_1| = 2$; $\alpha_{fs} = 0.7, \forall s \in \Omega_2$ with $|\Omega_2| = 2$; $\alpha_{fs} = 0.5, \forall s \in \Omega_3$ with $|\Omega_3| = 2$. Note that Ω_i is the subset uniformly drawn from the RRH set $\mathcal{S} = \{1, 2, 3, \ldots, S\}$. As such, we have $\cup \Omega_i = \Omega$. The value of α_{fs} is normalized by the standard deviation of the additive white Gaussian noise at RUE f, σ_f. Each curve is averaged over 50 different realizations. In each realization, the UE positions and the channel conditions vary. In Fig. 5, we plot the average number of admitted RUEs versus the target SINR of RUEs. As expected, the average number of admitted RUEs decreases as the target SINR of RUEs increases. However, it is seen that our proposed IRLS approach outperforms the MDR algorithm and its performance improves as p increases, where p is the parameter in l_p-norm. In Fig. 6, we plot the average transmission power consumption at the RRHs versus the target SINR of RUEs. It can be observed that the average transmission power consumption of our proposed IRLS approach is very close to that of the exhaustive search, indicating the effectiveness of our proposed UE selection algorithm.

Finally, the average number of admitted RUEs achieved by our proposed resource allocation scheme are compared with the baseline obtained from [33]. Note that, instead of minimizing the total sum-power as in [33], the FAP performs the resource allocation with the objective of maximizing the total tolerable interference threshold in our proposed scheme. It is seen that our proposed scheme can support a larger number of RUEs than the baseline for different values of target SINR of RUEs, implying that our proposed scheme can effectively increase the number of admitted RUEs in the F-RANs.

VII. Conclusion

In this paper, a framework of joint resource allocation and admission control was proposed for sliced F-RANs, in which NOMA technique and multicast beamforming are adopted in the IoT slice instance and the hotspot slice instance, respectively. In this framework, the the hotspot slice instance aims to maximize the number of RUEs that can be supported with desired QoS with the minimum transmission power, while satisfying the power budget and the interference threshold constraint for the IoT slice instance. To solve this non-convex problem, low-complexity algorithms were proposed to first determine the maximum total tolerable interference threshold for the FUEs via a joint sub-channel and power allocation, and then optimize the beamforming vector in the hotspot slice instance to maximize the number of RUEs with desired QoS. Through numerical results, it is shown that the proposed scheme for the IoT slice instance outperforms the conventional resource allocation scheme based on minimizing the power consumption, thus verifying its effectiveness.

There are still some other topics to be researched in the sliced F-RANs. It is interesting to extend our work to the larger diversity of use cases. For example, content caching will become more involved in a F-RAN that satisfies various service demands. The novel content caching strategies may consider unknown content popularity distribution and caching environment, joint allocation of caching and radio resource. Besides, consider the advanced modulation formats technique, F-RANs should be furthermore explored to improve the spectral efficiencies and to increase UE that can be supported with desired quality-of-service. The different performance indicators need to be considered to design the corresponding system model, such as latency and economical energy efficiency. The novel approaches may consider

machine learning based decision making and joint optimization of multi-dimensional resources.

Acknowledgement

This work was supported in part by the State Major Science and Technology Special Project (Grant No. 2018ZX03001002), in part by the National Natural Science Foundation of China under Grant No. 61925101 and No. 61831002, in part by the Beijing Natural Science Foundation under Grant No. JQ18016, in part by the National Program for Special Support of Eminent Professionals, and in part by the Fundamental Research Funds for the Central Universities under Grant No. 24820202020RC09 and Grant No.24820202020RC11.

References

[1] Z. Zhao, M. Xu, Y. Li and M. Peng, "A non-orthogonal multiple access-based multicast scheme in wireless content caching networks," *IEEE Journal Selected Areas in Communications*, vol. 35, no. 12, pp. 2723-2735, Dec. 2017.

[2] Z. Mao and S. Yan, "Deep learning based channel estimation in fog radio access networks," *China Communications*, vol. 16, no. 11, pp. 16-28, Nov. 2019.

[3] Z. Wang, F. Yang, S. Yan, S. Memon, Z. Zhao and C. Hu, "Joint design of coalition formation and semi-blind channel estimation in fog radio access networks," *China Communications*, vol. 16, no. 11, pp. 1-15, Nov. 2019.

[4] A. Kaloxylos, "A survey and an analysis of network slicing in 5G networks," *IEEE Communications Standards Magazine*, vol. 2, no. 1, pp. 60-65, Mar. 2018.

[5] H. Xiang, S. Yan and M. Peng, "A realization of Fog-RAN slicing via deep reinforcement learning," *IEEE Transactions on Wireless Communications*, vol. 19, no. 4, pp. 2515-2527, Apr. 2020.

[6] Y. Sun, M. Peng, S. Mao and S. Yan, "Hierarchical radio resource allocation for network slicing in fog radio access networks," *IEEE Transactions on Vehicular Technology*, vol. 68, no. 4, pp. 3866-3881, April 2019.

[7] R. F. Schaefer and H. Boche, "Physical layer service integration in wireless networks: signal processing challenges," *IEEE Signal Processing Magazine*, vol. 31, no. 3, pp. 147-156, May 2014.

[8] Y. Shi, J. Zhang and K. B. Letaief, "Robust group sparse beamforming for multicast green Cloud-RAN with imperfect CSI," *IEEE Transactions on Signal Processing*, vol. 63, no. 17, pp. 4647-4659, Sept.1, 2015.

[9] Z. Zhao, M. Peng, Z. Ding, W. Wang and H. V. Poor, "Cluster content caching: an energy-efficient approach to improve quality of service in cloud radio access networks," *IEEE Journal Selected Areas in Communications*, vol. 34, no. 5, pp. 1207-1221, May 2016.

[10] T. O. Olwal, K. Djouani and A. M. Kurien, "A survey of resource management toward 5G radio access networks," *IEEE Communications Surveys & Tutorials*, vol. 18, no. 3, pp. 1656-1686, third-quarter 2016.

[11] S. Yan, M. Peng, M. A. Abana and W. Wang, "An evolutionary game for user access mode selection in fog radio access networks," *IEEE Access*, vol. 5, pp. 2200-2210, 2017.

[12] Y. Sun, M. Peng and S. Mao, "Deep reinforcement learning based mode selection and resource management for green fog radio access networks," *IEEE Internet of Things Journal*, vol. 6, no. 2, pp. 1960-1971, April 2019.

[13] E. Matskani, N. D. Sidiropoulos, Z. Luo and L. Tassiulas, "Convex approximation techniques for joint multiuser downlink beamforming and admission control," *IEEE Transactions on Wireless Communications*, vol. 7, no. 7, pp. 2682-2693, July 2008.

[14] Y. Shi, J. Cheng, J. Zhang, B. Bai, W. Chen and K. B. Letaief, "Smoothed *lp*-minimization for green cloud-RAN with user admission control," *IEEE Journal Selected Areas in Communications*, vol. 34, no. 4, pp. 1022-1036, April 2016.

[15] S. Cical and V. Tralli, "QoS-Aware admission control and resource allocation for D2D communications underlaying cellular networks," *IEEE Transactions on Wireless Communications*, vol. 17, no. 8, pp. 5256-5269, Aug. 2018.

[16] 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 Journal Selected Areas in Communications*, vol. 35, no. 8, pp. 1809-1824, Aug. 2017.

[17] X. Sun, S. Yan, N. Yang, Z. Ding, C. Shen and Z. Zhong, "Short-packet downlink transmission with non-orthogonal multiple access," *IEEE Transactions on Wireless Communications*, vol. 17, no. 7, pp. 4550-4564, July 2018.

[18] Y. Sun, D. W. K. Ng, Z. Ding and R. Schober, "Optimal joint power and subcarrier allocation for full-duplex multicarrier non-orthogonal multiple access systems," *IEEE Transactions on Communications*, vol. 65, no. 3, pp. 1077-1091, March 2017.

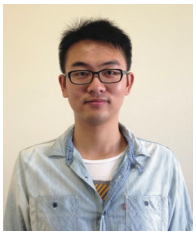
[19] J. Zhao, Y. Liu, K. K. Chai, A. Nallanathan, Y. Chen and Z. Han, "Spectrum allocation and power control for nonorthogonal multiple access in HetNets," *IEEE Transactions on Wireless Communications*, vol. 16, no. 9, pp. 5825-5837, Sept. 2017.

[20] J. Cui, Y. Liu, Z. Ding, P. Fan and A. Nallanathan, "QoE-based resource allocation for multi-cell NOMA networks," *IEEE Transactions on Wireless Communications*, vol. 17, no. 9, pp. 6160-6176, Sept. 2018.
[21] J. Liberti, S. Moshavi, and P. Zablocky, "Successive interference cancellation," U.S. Patent 8670418 B2, Mar. 11, 2014.
[22] A. Abdelnasser, E. Hossain and D. I. Kim, "Tier-aware resource allocation in OFDMA macrocell-small cell networks," *IEEE Transactions on Communications*, vol. 63, no. 3, pp. 695-710, March 2015.
[23] E. Bodine-Baron, C. Lee, A. Chong, B. Hassibi, and A. Wierman, "Peer effects and stability in matching markets," in *Proc. 4th Symp. Algorithmic Game Theory (SAGT)*, Amalfi, Italy, Oct. 2011, pp. 117-129.
[24] B. Di, L. Song and Y. Li, "Sub-channel assignment, power allocation, and user scheduling for non-orthogonal multiple access networks," *IEEE Transactions on Wireless Communications*, vol. 15, no. 11, pp. 7686-7698, Nov. 2016.
[25] S. Boyd and L. Vandenberghe, Convex Optimization. Cambridge, U.K.: Cambridge Univ. Press, 2004.
[26] E. Karipidis, N. D. Sidiropoulos, and Z.-Q. Luo, "Quality of service and max-min fair transmit beamforming to multiple cochannel multicast groups," *IEEE Transactions on Signal Processing*, vol. 56, pp. 1268C1279, Mar. 2008.
[27] D. Ge, X. Jiang, and Y. Ye, "A note on the complexity of minimization," *Math. Program.*, vol. 129, no. 2, pp. 285C299, 2011.
[28] D. Ba, B. Babadi, P. L. Purdon and E. N. Brown, "Convergence and stability of iteratively re-weighted least squares algorithms," *IEEE Transactions on Signal Processing*, vol. 62, no. 1, pp. 183-195, Jan.1, 2014.

Biographies

Yuan Ai, received the B.E. degree from University of Electronic Science and Technology of China(UESTC), Chengdu, China, in 2015. He is currently pursuing the Ph.D. degree in the State Key Laboratory of Networking and Switching Technology at Beijing University of Posts and Telecommunications (BUPT), Beijing, China. His research interest focuses on the non-orthogonal multiple access, resource management, and fog radio access networks.

Gang Qiu, Professorate senior engineer, Vice President of ZTE Corporation Wireless Product R&D Institute, mainly engaged in 4/5G project development and delivery. He has been in charge of national key science and technology projects and "863"(National High Technology Research and Development Program) projects for many times. Won the second prize of the National Science and Technology Progress Award, and one special prize of the National Science and Technology Progress Award. He owns more than 20 Chinese patents and 5 international patents.



Chenxi Liu, received his B.E. degree from Central South University, Changsha, China, in 2010, and Ph.D. degree from The University of New South Wales, Sydney, Australia, in 2016. From 2017 to 2019, he was a Postdoctoral research fellow in Singapore University of Technology and Design. Since 2019, he has been with the Beijing University of Posts and Telecommunications, where he is currently an Associate Professor. His research interests include wireless security, ultra-reliable and low-latency communications, internet-of-things, and network intelligence. He is currently serving as an Editor for the IEEE Wireless Communications Letters.

Yaohua Sun, received the bachelor's degree (with first class Hons.) in telecommunications engineering (with management) and Phd degree in communication engineering both from Beijing University of Posts and Telecommunications (BUPT), Beijing, China, in 2014 and 2019, respectively. He is currently an assistant professor at the State Key Laboratory of Networking and Switching Technology (SKL-NST), BUPT. His research interests include IoT, edge computing, resource management, (deep) reinforcement learning, network slicing, and fog radio access networks. He was the recipient of the National Scholarship in 2011 and 2017, and he has been reviewers for IEEE Transactions on Communications, IEEE Transactions on Mobile Computing, IEEE Systems Journal, Journal on Selected Areas in Communications, IEEE Communications Magazine, IEEE Wireless Communications Magazine, IEEE Wireless Communications Letters, IEEE Communications Letters, and IEEE Internet of Things Journal.