This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TVT.2019.2896586, IEEE Transactions on Vehicular Technology

Hierarchical Radio Resource Allocation for Network Slicing in Fog Radio Access Networks

Yaohua Sun, Mugen Peng, *Senior Member, IEEE*, Shiwen Mao, *Fellow, IEEE*, and Shi Yan

***Abstract*—Network slicing in fog radio access networks (F-RANs) is recognized as a cost-efficient solution to support future diverse use cases. However, with the number of user equipments (UEs) fast increasing, the centralized resource allocation architecture for network slicing can put heavy burdens on the global radio resource manager (GRRM), and meanwhile slice customization is not easy to achieve. To overcome the two issues, a hierarchical radio resource allocation architecture is proposed in this paper, where the GRRM is responsible for allocating subchannels to local radio resource managers (LRRMs) in slices, which then allocate the assigned resources to their UEs. Under this architecture, a hierarchical resource allocation problem is formulated, and the problem is further modeled as a Stakelberg game with the GRRM as the leader and LRRMs as followers, considering the hierarchy between the GRRM and LRRMs. Due to the NP-hardness of the followers' problems, a process based on exhaustive search is first proposed to achieve the Stackelberg equilibrium (SE). Nevertheless, when the network scale is large, achieving SE within limited decision making time is impractical for game players. Facing this challenge, the GRRM and LRRMs are seen as bounded rational players, and low complexity algorithms are developed to help them achieve local optimal solutions that lead to a weak version of SE. Simulation results show that there exists a tradeoff between the performance of slices, and the low complexity algorithms achieve close performance to that of exhaustive search and outperform other baselines significantly.**

***Index Terms*—Fog radio access networks (F-RANs), network slicing, radio resource allocation, Stackelberg game**

I. Introduction

To better satisfy future communication demands, fog radio access networks (F-RANs) are 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 [1]. In F-RANs, a user equipment (UE) can operate in different communication modes, including the centralized cloud-RAN mode and the fog access point (FAP) mode. Up to now, some interesting studies have been conducted on F-RANs, in terms of performance analysis [2], radio resource allocation [3], the joint design of cloud and edge processing [4], and so on.

On the other hand, as an emerging concept, network slicing is attracting more and more attentions, and is expected to be an efficient and flexible solution to support diverse application scenarios with different requirements. With network slicing, a physical network can be divided into multiple virtual networks, each of which is formed by a set of network functions and resources. When network slicing is conducted in an end-to-end manner, the RAN part of a network slice is referred to as a RAN slice, whose performance is dependent on resource allocation. In current works addressing resource allocation for RAN slicing, a critical aspect is to achieve slice isolation between RAN slices [5], and two different ways are often adopted to achieve this goal. The first one is allocating non-overlapping resources to different slices [7]–[13], while the second one is ensuring the minimum performance of each slice [14]. By isolation, any change in one slice will have no impact or at least a little impact on the performance of other slices.

Owing to the fog computing, cloud computing and heterogenous networking, F-RANs have great potentials in meeting diverse demands. For example, FAPs that are equipped with edge caches can be utilized to support the scenario where low latency is preferred, while the cloud RAN, which benefits from centralized signal processing and resource allocation [6], can be utilized in the scenario where high data rate is desired. Hence, in this paper, network slicing in F-RANs is considered, and resource allocation between two slices with different performance metrics is studied. Specifically, the aim of one slice is to minimize the content download latency, and the aim of the other one is to provide high data rate guarantee with minimal transmission power consumption. To alleviate the burdens on the global radio resource manager (GRRM) and achieve slice customization, a hierarchical radio resource allocation architecture is introduced, where the GRRM is responsible for allocating resources to each slice, and then the local radio resource manager (LRRM) in each slice allocates the assigned resources to its UEs. Under this architecture, a hierarchical resource allocation problem is formulated, which contains coupled subproblems, and the problem is further modeled as a Stackelberg game to produce a stable resource allocation outcome from which neither the GRRM nor the LRRM in each slice has incentive to deviate (i.e., to achieve the Stackelberg equilibrium (SE)).

Yaohua Sun (e-mail: sunyaohua@bupt.edu.cn) and Shi Yan (yanshi01@bupt.edu.cn) are with the Key Laboratory of Universal Wireless Communications (Ministry of Education), Beijing University of Posts and Telecommunications, Beijing, China. Mugen Peng (e-mail: pmg@bupt.edu.cn) is with the State Key Laboratory of Networking and Switching Technology (SKL-NST), Beijing University of Posts and Telecommunications, Beijing, China. Shiwen Mao (szm0001@auburn.edu) is with the Department of Electrical and Computer Engineering, Auburn University, Auburn, AL, USA. *(Corresponding author: Mugen Peng.)*

This work is supported in part by the State Major Science and Technology Special Project under 2017ZX03001025-006 and 2018ZX03001023-005, the National Natural Science Foundation of China under No. 61831002 and No. 61728101, the National Program for Special Support of Eminent Professionals, the National Science Foundation for Post-doctoral Scientists of China (Grant No. 2018M641279), and the US National Science Foundation under grant CNS-1702957.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TVT.2019.2896586, IEEE Transactions on Vehicular Technology

A. Related Work and Challenges

Recently, some efforts have been made in resource allocation for network slicing in wireless networks. In [9], all the transmitters, including a base station (BS) and device-to-device (D2D) transmitters, are sliced by a hypervisor to serve users. Considering the imperfect channel state information (CSI) observed by the hypervisor, the joint transmitter association, subchannel allocation and caching optimization is formulated as a discrete stochastic optimization problem aiming at maximizing the network utility, which is efficiently solved with discrete stochastic approximation. In [10], a mobile virtual network operator (MVNO) rents physical resources from multiple infrastructure providers (InPs) and divides these resources into multiple slices, each of which is composed of a BS and a subcarrier chunk. In [14], a downlink orthogonal frequency division multiple access based wireless network is considered, where multiple slices with multiple users coexist. To maximize the network sum rate, a joint BS assignment, subcarrier allocation and power allocation problem is investigated, where the impact of interference between slices is limited by guaranteeing the minimum user sum rate of each slice. To deal with this non-convex mixed-integer problem, successive convex approximation and complementary geometric programming are applied.

Although the proposed approaches in [9], [10], [14] can achieve good performance, allocating resources directly to all the users can put heavy burdens on the central controller, especially when the number of users is large. Meanwhile, by such centralized resource allocation, the customization of each slice, which means each slice can individually decide how to allocate available resources, cannot be easily achieved. To overcome these two issues, a hierarchical resource allocation architecture is adopted in [8], [13], where the resources are firstly allocated to each slice by a central controller, and then each slice allocates the assigned resources to its own users. Specifically, in [8], the resource allocation among slices is formulated as a hierarchical auction game, where slice isolation is achieved by allocating non-overlapping resources to slices. In the upper layer, the InP, acting as the seller as well as the auctioneer, divides the owned resources into multiple bundles, each of which includes subchannels, power and antennas, and each MVNO acts as a buyer who submits a bid for each resource bundle. In the lower layer, each MVNO acts as a seller with users acting as buyers. Similar to [8], an auction game is used to model the resource allocation to slices, where slices have the incentive to bid truthfully to compete for resource blocks.

In [15], the authors summarize two cases of using slices. The first one is called Quality of Service Slicing, which is to support different services by creating dedicated slices, while the other is termed Infrastructure Sharing Slicing, which focuses on infrastructure sharing among multi-tenants. Unfortunately, the works [8], [13], together with [9]–[12], [14], all fail to explicitly consider the first case, which is envisioned as a key characteristic of the fifth generation mobile network. In addition, in works [8], [10]–[14], network slicing is done in the RANs that cannot well support diverse scenarios. For example, the cellular networks with massive antennas considered in [8], [11] can well support hot spot areas, but the download latency can be large since the contents requested by the users need to be fetched from a remote content server. Therefore, network slicing in more advanced RANs should be investigated. To this end, network slicing is done in a downlink F-RAN in this paper. Specifically, one slice is a traditional C-RAN to provide high data rate, and the other slice is composed of FAPs with edge caching capabilities to provide low download latency. Under a hierarchical resource allocation architecture, resource allocation among slices is modeled as a Stackelberg game with the GRRM as the leader and LRRMs as followers, in which the GRRM first allocates a set of subchannels to each LRRM and then each LRMM decides the resource allocation within its slice.

Previously, game theory has been widely applied to resource allocation problems in heterogenous networks [16]–[27]. In [16]–[19], coalitional game is adopted to model the cooperative behavior among BSs or users that have equal position, but this kind of game is not suitable for our problem, since there is hierarchy between the GRRM and LRRMs in terms of decision making. In [20]–[23], authors utilize Stackelberg game to design pricing schemes to control inter-tier interference between the leader/leaders and the follower/followers, which is not the focus of this paper. While in [24], [25], the leader, which is the macro BS in [24] and the relay node in [25], only cares about its own resource allocation. However, in our considered network slicing scenario, the leader, which is the GRRM, needs to optimize the resource allocation to each follower. In [26], the cloud as the leader directly allocates a set of serving nodes to each user, and the leader in [27], which is a cognitive BS, directly allocates the spectrum bought from primary networks to each user served by a femto BS. On the contrary, in our work, the GRRM as the leader does not allocate resources to each user directly. Instead, to meet the requirement of slice customization, the GRRM first decides the resource allocation between slices, and then the resource allocation to each user is optimized by the LRRM in each slice. In addition, different from [20]–[27]. where the follower is a BS or a user, each follower in our problem is a resource manager of a slice consisting of multiple BSs and multiple users, and the strategy of each follower is the resource allocation for the whole slice. This unavoidably makes deriving the optimal strategy of each follower and SE more challenging. Particularly, the problem related to one LRRM is a mixed-integer nonlinear programming that is NP-hard [28], [29] and the problem associated with the other LRRM is a non-convex problem with 0-1 variables that is also NP-hard [30].

B. Contributions and Organization

This paper proposes a Stakelberg game based approach to radio resource allocation for network slicing in a downlink F-RAN to address the challenges incurred by slice customization and heterogenous slice performance requirements, which can achieve a stable allocation result. In the considered scenario, there exist two slices, namely slice $s1$ and $s2$. Specifically,

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TVT.2019.2896586, IEEE Transactions on Vehicular Technology

slice $s1$ aims to provide guaranteed high data rates with minimal transmission power consumption by centralized signal processing and resource allocation in the cloud, while slice $s2$ takes advantage of the edge caching capabilities of FAPs to minimize the content download latency. The main contributions of the paper are:

- To alleviate the burdens on the GRRM and achieve slice customization, a hierarchical radio resource allocation architecture is employed for network slicing in F-RANs. The GRRM is responsible for allocating radio resources to each slice, and then the LRRM in each slice allocates the assigned resources to its UEs. Under this architecture, a hierarchical resource allocation problem is formulated containing coupled subproblems, each of which relates to the GRRM or an LRRM. The objectives of LRRM 1 in slice $s1$ and LRRM 2 in slice $s2$ are to provide high data rate with minimal transmission power consumption and to minimize the download latency, respectively, while the GRRM aims to optimize the performance of slice $s1$ meanwhile guaranteeing the performance of slice $s2$.
- Considering the hierarchical decision making between the GRRM and LRRMs and the coupling of their strategies, the problem is further modeled as a Stakelberg game, where the GRRM serves as the leader and the two LRRMs are taken as followers. Moreover, the stable state of the game, i.e., SE, is defined, and the existence as well as the uniqueness of the equilibrium are analyzed.
- Different from the Stackelberg game formulated in previous works where the follower is a BS or a user, each follower in our work is a resource manager of a slice consisting of multiple BSs and multiple users, which makes deriving the optimal strategy of each follower very challenging. Particularly, the problem related to LRRM 1 is an MINLP and the problem associated with LRRM 2 is a non-convex problem with 0-1 variables. Moreover, the strategy of the leader is discrete. Faced with the two issues, a process based on exhaustive search is introduced to achieve SE when the players are completely rational, and meanwhile low complexity algorithms are proposed as well for the case where the GRRM and LRRMs are bounded rational. Furthermore, the complexity and optimality of all the proposals are rigorously analyzed and simulation is conducted to demonstrate their effectiveness.

The remainder of this paper is organized as follows. Section II describes the downlink F-RAN with two slices. Section III formulates the related optimization problem, and Section IV presents a Stackelberg game based approach to find SE. Section V develops low complexity resource allocation algorithms for bounded rational GRRM and LRRMs, and simulation results are presented in Section VI, followed by the conclusions in Section VII.

II. System Model

In this section, the model of a downlink F-RAN is firstly provided. We then present the models for the two slices.

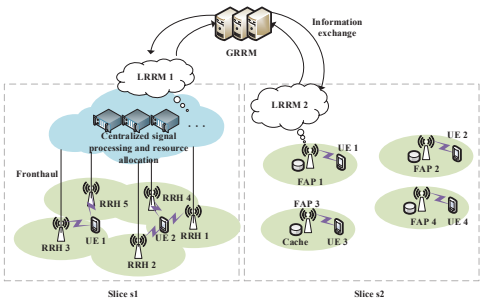

Fig. 1. Network slicing in a downlink F-RAN with a hierarchical resource allocation architecture.

A. Network Slicing in a Downlink F-RAN

The downlink F-RAN shown in Fig. 1 consists of one cloud, multiple remote radio heads (RRHs), multiple FAPs and multiple D2D transmitters. The F-RAN is divided into two slices, $s1$ and $s2$, which intend to provide high data rate guarantee and low download latency, respectively. Note that high data rate can be achieved by taking advantage of centralized signal processing and resource allocation in the cloud, while content downloading latency can be reduced by utilizing the caching capabilities of FAPs. Hence, slice $s1$ is allocated with a cloud and K RRHs whose set is denoted by $\mathcal{K} = \{1, 2, ..., K\}$, with each RRH equipped with N antennas. Slice $s2$ is allocated with L single-antenna FAPs, whose set is denoted as $\mathcal{L} = \{1, 2, ..., L\}$. It should be highlighted that although flexible slice configuration is supported by network slicing, infrastructure configuration is usually done over a large time scale [31], which should take the slice performance metrics and the budget of customers into account, while subchannel allocation is performed over a smaller time scale to adapt to the radio environment. In this paper, subchannel allocation for network slicing is studied, assuming that the infrastructure configuration of slices is fixed and cooperative transmission is only applied in slice $s1$. Denote the set of UEs with single antenna served by slice $s1$ and $s2$ as $\mathcal{M}_{s1} = \{1, 2, ..., M_{s1}\}$ and $\mathcal{M}_{s2} = \{1, 2, ..., M_{s2}\}$, respectively. The set of available subchannels in the system is denoted by $\mathcal{D} = \{1, 2, ..., D\}$, and each of them is with bandwidth B.

To alleviate the burden on the GRRM and achieve slice customization, a hierarchical radio resource allocation architecture is adopted, which consists of a GRRM and two LRRMs. Specifically, the GRRM is responsible for allocating subchannels to slices based on only the performance feedback from LRRMs and some coarse information about slices, while the LRRM in each slice is responsible for allocating the assigned resources to its UEs. Under this setting, the resource allocation solution space of the GRRM is greatly reduced, which does not depend on the network scale in each slice. In addition, the GRRM does not need to gather global CSI and each LRRM can determine its own resource allocation strategy. In the following, the LRRM in slice $s1$ and the LRRM in slice $s2$ are named as LRRM 1 and LRRM 2, respectively. For the sake of slice isolation, the GRRM allocates slice $s1$ and $s2$ with disjoint sets of subchannels denoted by $\mathcal{D}_{s1}$ and

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TVT.2019.2896586, IEEE Transactions on Vehicular Technology

$\mathcal{D}_{s2}$, respectively. Note that the GRRM and LRRMs can be implemented in software running on computing platforms. For example, LRRM 1 can be located in the powerful servers in the cloud, while LRRM 2 can be deployed at an FAP with high edge computing capability and backhaul links of good quality that help collect system information from other connected FAPs. Finally, the GRRM can be run in the slice orchestration entity defined in [32].

B. The Slice $s1$ Model

In slice $s1$, each UE is assumed to be allocated with a subchannel, and each subchannel can be reused by multiple UEs. For UE $m_{s1,j} \in \mathcal{M}_{s1}$, its received signal is written as

$$c_{m_{s1,j},d} = \sum_{k \in \mathcal{K}_{m_{s1,j}}} \mathbf{h}^H_{m_{s1,j},k,d} \mathbf{v}_{m_{s1,j},k,d} x_{m_{s1,j}} + \sum_{m_{s1,i} \in \mathcal{M}_{s1}, i \neq j, \sum_k w_{m_{s1,i},k,d} > 0} \sum_{k \in \mathcal{K}_{m_{s1,i}}} \mathbf{h}^H_{m_{s1,j},k,d} \mathbf{v}_{m_{s1,i},k,d} x_{m_{s1,i}} + z_{m_{s1,j},d}, \tag{1}$$

where $x_{m_{s1,i}}$ is the message of UE $m_{s1,i}$, $w_{m_{s1,i},k,d}$ is a 0-1 indicator which equals to 1 when RRH k serves UE $m_{s1,i}$ over subchannel d, $\mathcal{K}_{m_{s1,i}}$ is the set of RRHs satisfying $w_{m_{s1,i},k,d} = 1$, i.e., the set of RRHs serving UE $m_{s1,i}$. $\sum_k w_{m_{s1,i},k,d} > 0$ means that UE $m_{s1,i}$ is allocated with subchannel d. Furthermore, $\mathbf{h}_{m_{s1,j},k,d}$ is the channel vector between RRH k and UE $m_{s1,j}$ on subchannel d, $\mathbf{v}_{m_{s1,j},k,d}$ is the precoding vector of RRH k for UE $m_{s1,j}$, and $z_{m_{s1,j},d}$ is the noise which follows the distribution of $\mathcal{CN}\left(0, \sigma^2\right)$. Then the data rate of UE $m_{s1,j}$ is calculated by equation (2).

C. The Slice $s2$ Model

For slice $s2$, it is assumed that each UE is allocated with one subchannel, but each subchannel can be reused by multiple UEs accessing different FAPs. Suppose that each UE $m_{s2,j}$ in slice $s2$ randomly requests a content $f_{m_{s2,j}}$, and each FAP selects popular contents to cache. If desired contents are cached at associated FAPs, UEs will be served locally avoiding the latency induced by fetching contents from the cloud. For UE $m_{s2,j}$, when it is served by an FAP l over subchannel d, its received signal is written as

$$c_{m_{s2,j},l,d} = p_l h_{m_{s2,j},l,d} x_{m_{s2,j}} + \sum_{m_{s2,i} \in \mathcal{M}_{S2,d}, i \neq j} \sqrt{p_{l_{m_{s2,i}}}} h_{m_{s2,j},l_{m_{s2,i}},d} x_{m_{s2,i}} + z_{m_{s2,j},d}, \tag{3}$$

where p_l is the transmit power of FAP l on each subchannel, which is assumed to be a constant value, $h_{m_{s2,j},l,d}$ is the channel gain between UE $m_{s2,j}$ and FAP l over subchannel d, and $l_{m_{s2,i}}$ denotes the FAP serving UE $m_{s2,i}$. Hence, the data rate of UE $m_{s2,j}$ is given by

$$R_{m_{s2,j},l,d} = B \log \left(1 + \frac{p_l \left| h_{m_{s2,j},l,d} \right|^2}{\sigma^2 + \sum_{m_{s2,i} \in \mathcal{M}_{S2,d}, i \neq j} p_{l_{m_{s2,i}}} \left| h_{m_{s2,j},l_{m_{s2,i}},d} \right|^2} \right). \tag{4}$$

By involving a 0-1 indicator $y_{m_{s2,j},l,d}$, which equals to 1 if and only if UE $m_{s2,j}$ is served by FAP l over subchannel d, the data rate of UE $m_{s2,j}$ can be expressed as

$$R_{m_{s2,j}}\left(\mathcal{D}_{s2}, \left\{y_{m_{s2,j},l,d}\right\}\right) = \sum_{d \in \mathcal{D}_{s2}} \sum_{l \in \mathcal{L}} y_{m_{s2,j},l,d} R_{m_{s2,j},l,d}. \tag{5}$$

Then, the content download latency of UE $m_{s2,j}$ is given by

$$t_{m_{s2,j}}\left(\mathcal{D}_{s2}, \left\{y_{m_{s2,j},l,d}\right\}\right) = \begin{cases} \frac{s_{m_{s2,j}}}{R_{m_{s2,j}}}, when\ f_{m_{s2,j}}\ is\ cached\ at\ FAP\ l\ serving\ UE\ m_{s2,j}, \\ \frac{s_{m_{s2,j}}}{R_{m_{s2,j}}} + \beta_{m_{s2,j}},\ otherwise, \end{cases} \tag{6}$$

where $s_{m_{s2,j}}$ is the size of the content requested by UE $m_{s2,j}$, and $\beta_{m_{s2,j}}$ is the latency needed for an FAP to download the requested file from the cloud via a fronthaul link.

III. Problem Formulation

In this section, under the hierarchical radio resource allocation architecture, a radio resource allocation problem containing three subproblems is formulated, each of which relates to the GRRM or an LRRM.

A. Problem Formulation for LRRM 1

By centralized signal processing and radio resource allocation in the cloud, the aim of slice $s1$ is to provide high data rate guarantee with minimum transmission power consumption. The optimization problem for LRRM 1 is formulated as follows:

$$\begin{aligned} &\min_{\left\{w_{m_{s1,j},k,d}\right\}, \left\{\mathbf{v}_{m_{s1,j},k,d}\right\}} U_1 \\ &(a1)\ \left[\sum_k w_{m_{s1,j},k,d}\right] \left[\sum_{d' \neq d} \sum_k w_{m_{s1,j},k,d'}\right] = 0, \forall m_{s1,j}, \forall d, \\ &(a2)\ \sum_d \sum_k w_{m_{s1,j},k,d} > 0, \forall m_{s1,j}, \\ &(a3)\ R_{m_{s1,j}}\left(\mathcal{D}_{s1}, \left\{w_{m_{s1,j},k,d}\right\}, \left\{\mathbf{v}_{m_{s1,j},k,d}\right\}\right) \geq R_{m_{s1,j},\min}, \\ &\forall m_{s1,j}, \\ &(a4) \sum_{m_{s1,j}} \sum_{d \in \mathcal{D}_{s1}} w_{m_{s1,j},k,d} \leq o_k, \forall k, \\ &(a5) \sum_{m_{s1,j}} \sum_{d \in \mathcal{D}_{s1}} w_{m_{s1,j},k,d} \left\|\mathbf{v}_{m_{s1,j},k,d}\right\|_2^2 \leq p_{\max}, \end{aligned} \tag{7}$$

where $U_1 = \sum_k \sum_{m_{s1,j}} \sum_{d \in \mathcal{D}_{s1}} w_{m_{s1,j},k,d} \left\|\mathbf{v}_{m_{s1,j},k,d}\right\|_2^2$. The first constraint states that once UE $m_{s1,j}$ is allocated with subchannel d, it will not be allocated with other subchannels. The second constraint ensures that each UE must be allocated with at least one subchannel, which, together with the first constraint, jointly states that each UE is allocated with one

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TVT.2019.2896586, IEEE Transactions on Vehicular Technology

$$R_{m_{s1,j}}\left(\mathcal{D}_{s1},\left\{w_{m_{s1,j},k,d}\right\},\left\{\mathbf{v}_{m_{s1,j},k,d}\right\}\right)= \sum_{d\in\mathcal{D}_{s1},\sum\limits_{k} w_{m_{s1,j},k,d}>0} B\log\left(1+\frac{\left|\sum\limits_{k\in\mathcal{K}_{m_{s1,j}}}\mathbf{h}^{H}_{m_{s1,j},k,d}\mathbf{v}_{m_{s1,j},k,d}\right|^{2}}{\sum\limits_{m_{s1,i}\in\mathcal{M}_{s1},i\neq j,\sum\limits_{k} w_{m_{s1,i},k,d}>0}\left|\sum\limits_{k\in\mathcal{K}_{m_{s1,i}}}\mathbf{h}^{H}_{m_{s1,j},k,d}\mathbf{v}_{m_{s1,i},k,d}\right|^{2}+\sigma^{2}}\right). \tag{2}$$

subchannel. The third constraint is to provide high data rate guarantee. The fourth constraint is the fronthaul capacity constraint, which is enforced by limiting the maximal number of UEs that can be supported by a fronthaul link [33]. The last constraint is the transmit power constraint per RRH.

B. Problem Formulation for LRRM 2

Aiming to minimize the content download latency, the optimization problem for LRRM 2 is given by

$$\begin{aligned}&\min_{\left\{y_{m_{s2,j},l,d}\right\}} U_2=\tfrac{1}{M_{s2}}\sum_{m_{s2,j}} t_{m_{s2,j}}\left(\mathcal{D}_{s2},\left\{y_{m_{s2,j},l,d}\right\}\right)\\ &(b1)\ \sum_{d\in\mathcal{D}_{s2}}\sum_{l} y_{m_{s2,j},l,d}=1,\forall m_{s2,j},\\ &(b2)\ \sum_{m_{s2,j}} y_{m_{s2,j},l,d}\leq 1,\forall 1\leq l\leq L,\ \forall d\in\mathcal{D}_{s2},\\ &(b3)\ \sum_{d\in\mathcal{D}_{s2}}\sum_{m_{s2,j}} y_{m_{s2,j},l,d}\leq q,\forall 1\leq l\leq L.\end{aligned} \tag{8}$$

The first constraint requires that each UE can be allocated with only one FAP and one subchannel. The second constraint guarantees that the UEs served by an FAP occupy orthogonal subchannels, and the third constraint together with constraint $(b2)$ states that each FAP can serve at most q UEs due to the total transmission power constraints. Note that the second constraint implicity indicates that the maximal number of UEs that can be served by an FAP is also limited by the number of available subchannels of slice $s2$, i.e., $|\mathcal{D}_{s2}|$, and $t_{m_{s2,j}}$ is affected by the content placement according to (6). Denote the set of contents potentially requested by UEs as $\mathcal{F}=\{1,2,...,F\}$ and define a content placement matrix $\mathbf{\Omega}$ with the size of $L\times F$ where its element $\Omega_{l,f}$ represents whether FAP l caches content f. In addition, the contents in $\mathcal{F}$ follow a certain popularity distribution like Zipf distribution, and the cache size of each FAP is Γ. Finally, performance metric in (8) can be calculated when contents have been cached at each FAP by caching the most popular Γ contents, and optimizing radio resource under fixed content placement is reasonable, since content placement is generally adjusted on a large time scale [31].

C. Problem Formulation for the GRRM

As a global controller, the GRRM should take into account the performance of both slices. However, since the performance goals of the two slices can have conflict, minimizing both of them may not be realistic sometimes. Therefore, it is assumed that the GRRM aims to minimize the transmit power consumption of slice $s1$, while guaranteeing the download latency performance of slice $s2$. The optimization problem is formulated as follows:

$$\begin{aligned}&\min_{\mathcal{D}_{s1},\mathcal{D}_{s2}} U_0=\begin{cases}U_1,\ if\ U_2\leq T_{\max}\\ Kp_{\max},\ otherwise\end{cases}\\ &(c1)\ \mathcal{D}_{s1}\cup\mathcal{D}_{s2}=\mathcal{D},\\ &(c2)\ \mathcal{D}_{s1}\cap\mathcal{D}_{s2}=\phi,\\ &(c3)\ \lceil\tfrac{M_{s2}}{L}\rceil\leq|\mathcal{D}_{s2}|\leq M_{s2}.\end{aligned} \tag{9}$$

The first constraint means the GRRM will allocate all the available subchannels to the two slices. The second constraint is used to avoid the interference between slices and hence slice isolation is guaranteed. The third constraint is to ensure that there is enough number of subchannels for the UEs in slice $s2$ considering the different degrees of subchannel reuse, where $\lceil\rceil$ represents the ceiling function. Moreover, T_{max} is the maximal tolerable latency of slice $s2$.

In the definition of the GRRM utility, the utility value is set to the maximal transmission power consumption of slice $s1$ as a punishment when the performance requirement of slice $s2$ is not met, which facilitates the GRRM to take the performances of both slices into account. Note that in the applications of game theory to wireless networks, the utilities of players are not always completely conflicting. For example, the utilities of players in a non-cooperative game can be identical and correspond to the global performance [34], and the cloud, as the leader in the concerned Stackelberg game in [27], intends to maximize the sum rate of all the users that are followers. Hence, the utility selection of the GRRM in our paper is reasonable.

At the end of this section, we highlight the difficulties of our work compared to literatures studying resource allocation in multi-tier heterogenous networks (HetNets), and note that each slice in our work can be seen as a tier in a HetNet. In [35], a distributed power adjustment algorithm is designed, and power allocation for a multi-tier network is also studied in [36]. However, the optimization of user-BS association and subchannel allocation is not considered in [35] and [36], which needs to be addressed in our work for both tiers. In [37], the bias factor of each tier for user association and the proportion of spectrum allocated to each tier are optimized based on stochastic geometry analysis. Nevertheless, stochastic geometry analysis can not help us to get exact resource allocation of each tier and each user based on instantaneous CSI. In [38] and [39], resource is directly allocated to each user. In our work, in addition to allocating resource to each user, which is performed by LRRMs, we also have to address the resource allocation between the considered two tiers, which has significant impacts on the user resource allocation

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TVT.2019.2896586, IEEE Transactions on Vehicular Technology

result. In [40], bandwidth resource allocation refers to that each wireless service provider needs to decide the amount of bandwidth allocated to each tier of the heterogenous network it owns. However, the optimization of resource allocation to each user within each tier is lacked, which needs to be handled by LRRMs in this paper via solving an NP-hard resource allocation problem for each tier.

IV. A Stackelberg Game Approach

In this section, the formulated resource allocation problem is modeled as a Stackelberg game, and the existence and uniqueness of the SE are analyzed. In addition, a method based on exhaustive search is proposed to identify the existence of SE and find SE (if SE exists) at the same time.

A. Stackelberg Game Formulation

The resource allocation problem formulated in the previous section is a hierarchical optimization problem, where the GRRM optimization problem is the upper level problem while the LRRM optimization problems are lower level problems. The upper level problem and lower level problems are tightly coupled. Note that under the hierarchical resource allocation architecture, the GRRM holds a strong position and each LRRM can only react to its allocation result. Hence, inspired by [41], the resource allocation problem is thereby naturally formulated as a Stackelberg game taking the GRRM as the leader and LRRMs as followers, where the equilibrium state of the game, i.e., SE, can be adopted as the resource allocation outcome from which neither the GRRM nor the LRRM in each slice is willing to deviate for improved utility.

The strategy of the GRRM with utility U_0 is the resource allocation among the two slices, i.e., $\mathcal{D}_{s1}$ and $\mathcal{D}_{s2}$, which can be further denoted by a $1 \times D$ subchannel allocation vector $\mathbf{d}$ whose elements are 0-1 variables. Specifically, $d_i = 1$ means subchannel i is allocated to LRRM 1, and $d_i = 0$ means subchannel i is allocated to LRRM 2. The strategy of LRRM 1 with utility U_1 is the resource allocation among the UEs in slice $s1$, i.e., $\left\{w_{m_{s1,j},k,d}\right\}$ and $\left\{\mathbf{v}_{m_{s1,j},k,d}\right\}$. $\left\{w_{m_{s1,j},k,d}\right\}$ can be further denoted by a $1 \times M_{s1} K \left|\mathcal{D}_{s1}\right|$ vector $\mathbf{w}$, where each element relates to a $w_{m_{s1,j},k,d}$. The strategy of LRRM 2 with utility U_2 is the resource allocation among the UEs in slice $s2$, i.e., $\left\{y_{m_{s2,j},l,d}\right\}$, which is further denoted by a $1 \times M_{s2} L \left|\mathcal{D}_{s2}\right|$ vector $\mathbf{y}$, where each element relates to a $y_{m_{s2,j},l,d}$.

After clarifying the strategies and utilities of the players, the stable state of the formulated game is given by the following definition that follows reference [41].

Definition *1:* The strategy $\mathbf{d}^*$ is an SE strategy for the leader, i.e., the GRRM, if

$$\begin{aligned} &U_0\left(\mathbf{d}^*, \left\{\mathbf{w}^*, \left\{\mathbf{v}^*_{m_{s1,j},k,d}\right\}\right\}_{\mathbf{d}^*}, \mathbf{y}^*_{\mathbf{d}^*}\right) \\ &= \min_{\mathbf{d}} U_0\left(\mathbf{d}, \left\{\mathbf{w}^*, \left\{\mathbf{v}^*_{m_{s1,j},k,d}\right\}\right\}_{\mathbf{d}}, \mathbf{y}^*_{\mathbf{d}}\right), \end{aligned} \tag{10}$$

where $\left\{\mathbf{w}^*, \left\{\mathbf{v}^*_{m_{s1,j},k,d}\right\}\right\}_{\mathbf{d}}$ and $\mathbf{y}^*_{\mathbf{d}}$ are the optimal strategies of slice $s1$ and $s2$ reacting to the leader's strategy $\mathbf{d}$, respectively. If there is such a $\mathbf{d}^*$, $\mathbf{d}^*$, together with $\left\{\mathbf{w}^*, \left\{\mathbf{v}^*_{m_{s1,j},k,d}\right\}\right\}_{\mathbf{d}^*}$ and $\mathbf{y}^*_{\mathbf{d}^*}$, constitutes an SE solution to our considered hierarchical resource allocation game.

Although taking SE as the resource allocation outcome can inevitably result in the inefficiency. However, more efficient algorithms like a fully centralized approach, which can further improve the utilities of all the players, may lose several benefits, such as reducing the computation burden on the GRRM, avoiding the collection of CSI within each slice and easier intra-slice customization.

Next, the existence and uniqueness of the equilibrium are presented in the following theorem.

Theorem *1:* If there exists a resource allocation strategy $\mathbf{d}$ of the GRRM, under which the constraints in (7), (8) and (9) hold at the same time, the Stackelberg game must have at least one SE.

Proof: When there does not exist a resource allocation strategy $\mathbf{d}$ of the GRRM under which the constraints in (7), (8) and (9) are satisfied simultaneously, there is no feasible $\mathbf{d}$ at all and hence no SE exists. On the contrary, we can always find a feasible d satisfying (10) in Definition 1, and hence SE must exist. To identify the existence of SE and find SE (if SE exists) at the same time, Algorithm 1 is presented in the next subsection. Specifically, at the beginning of Algorithm 1, the GRRM first generates a set $\mathcal{D}'$ of $\mathbf{d}$ meeting $(c1)$, $(c2)$ and $(c3)$ in (9). If $\mathcal{D}'$ is empty, Algorithm 1 terminates and there is no SE solution. If $\mathcal{D}'$ is non-empty, the GRRM starts to try each $\mathbf{d}$ in $\mathcal{D}'$. Once receiving a GRRM's strategy $\mathbf{d}$, each LRRM feeds back its optimal utility value to the GRRM if its problem under $\mathbf{d}$ is feasible. Otherwise, each LRRM feeds back a signaling indicating problem infeasibility. After receiving feebacks from both LRRMs, the GRRM adds current $\mathbf{d}$ to a set $\mathcal{Q}$ and records its achieved utility $U_0(\mathbf{d})$ only if both LRRMs feed back their utility values. The process continues until all the $\mathbf{d}$ in $\mathcal{D}'$ have been searched. Finally, if $\mathcal{Q}$ is empty, there is no SE solution. Otherwise, the SE strategy is derived as

$$\mathbf{d}^* = \arg\min_{\mathbf{d} \in \mathcal{Q}} U_0^*(\mathbf{d}), \tag{11}$$

which, together with the corresponding optimal strategies of both LRRMs under $\mathbf{d}^*$, constitutes the SE of the game.

However, the uniqueness of the SE is not guaranteed, since LRRM 1 not necessarily allocates all the assigned subchannels to its UEs and hence it is possible for two different subchannel allocation vectors $\mathbf{d_1}$ and $\mathbf{d_2}$ to lead to the same U_0. When there exist multiple SEs, the SE with the best performance for LRRM 2 is selected to improve the efficiency of resource allocation. ■

B. The Solution of the Game

In the Stackelberg game, rational followers react optimally to the strategy of the leader. Here, finding the optimal strategies of LRRM 1 and 2 are equivalent to solving problem (7) and (8) optimally under a given $\mathbf{d}$, respectively. In the following, we first analyze problem (7) of LRRM 1.

When a $\mathbf{w}$ satisfying $(a1)$, $(a2)$ and $(a4)$ is given, problem (7) can be simplified to the following precoding design

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TVT.2019.2896586, IEEE Transactions on Vehicular Technology

problem:

$$\begin{aligned}&\min_{\left\{\mathbf{v}_{m_{s1,j},k,d_{m_{s1,j}}}\right\}} \sum_{m_{s1,j}\in\mathcal{M}_{s1}} \sum_{k} \left\|\mathbf{v}_{m_{s1,j},k,d_{m_{s1,j}}}\right\|_2^2\\ &(d1)\ R_{m_{s1,j}} \geq R_{m_{s1,j},\min}, \forall m_{s1,j},\\ &(d2) \sum_{m_{s1,j}} \left\|\mathbf{v}_{m_{s1,j},k,d_{m_{s1,j}}}\right\|_2^2 \leq p_{\max}, \forall k,\\ &(d3)\ \mathbf{v}_{m_{s1,j},k,d_{m_{s1,j}}} = \mathbf{0},\ if\ w_{m_{s1,j},k,d_{m_{s1,j}}} = 0,\end{aligned} \tag{12}$$

where $d_{m_{s1,j}}$ denotes the subchannel allocated to UE $m_{s1,j}$, and constraint $(d3)$ states that if UE $m_{s1,j}$ is not associated with RRH k, the corresponding precoding vector is set to a zero vector. Note that constraint $(d1)$ can be transformed into a convex second order cone constraint by rotating the phase of $\mathbf{v}_{m_{s1,j},k,d_{m_{s1,j}}}$, which will not influence the objective and the constraints. Then, problem (12) can be further rewritten as

$$\begin{aligned}&\min_{\left\{\mathbf{v}_{m_{s1,j},k,d_{m_{s1,j}}}\right\}} \sum_{m_{s1,j}\in\mathcal{M}_{s1}} \sum_{k} \left\|\mathbf{v}_{m_{s1,j},k,d_{m_{s1,j}}}\right\|_2^2\\ &(e1)\ \sqrt{\sum_{m_{s1,i},d_{m_{s1,i}}=d_{m_{s1,j}}} \left|\mathbf{h}_{m_{s1,j},d_{m_{s1,j}}}^H \mathbf{v}_{m_{s1,i}}\right|^2 + \sigma^2}\\ &\leq \sqrt{1+\frac{1}{\gamma_{m_{s1,j}}}} Re\left\{\mathbf{h}_{m_{s1,j},d_{m_{s1,j}}}^H \mathbf{v}_{m_{s1,j}}\right\}, \forall m_{s1,j},\\ &(e2)\ \mathrm{Im}\left\{\mathbf{h}_{m_{s1,j},d_{m_{s1,j}}}^H \mathbf{v}_{m_{s1,j}}\right\} = 0, \forall m_{s1,j},\\ &(e3) \sum_{m_{s1,j}} \left\|\mathbf{v}_{m_{s1,j},k,d_{m_{s1,j}}}\right\|_2^2 \leq p_{\max}, \forall k,\\ &(e4)\ \mathbf{v}_{m_{s1,j},k,d_{m_{s1,j}}} = \mathbf{0},\ if\ w_{m_{s1,j},k,d_{m_{s1,j}}} = 0,\end{aligned} \tag{13}$$

where $\gamma_{m_{s1,j}} = 2^{\frac{R_{m_{s1,j},\min}}{B}} - 1$ is the minimum requirement of the signal to interference plus noise ratio (SINR) of UE $m_{s1,j}$, $\mathbf{h}_{m_{s1,j},d_{m_{s1,j}}}$ is the channel vector from all the RRHs to UE $m_{s1,j}$ on subchannel $d_{m_{s1,j}}$, and $\mathbf{v}_{m_{s1,j}}$ is the network wide precoding vector of UE $m_{s1,j}$.

The above problem is convex that can be efficiently solved by CVX. Then, the optimal strategy of LRRM 1 under a given $\mathbf{d}$ can be got by an exhaustive search as follows. First, generate all the possible $\mathbf{w}$ that meet constraints $(a1)$, $(a2)$ and $(a4)$. Second, calculate the optimal power consumption by solving problem (13) for each $\mathbf{w}$. Then, the $\mathbf{w}$ leading to the minimum power consumption, together with the corresponding optimal precoding vector $\left\{\mathbf{v}_{m_{s1,j},k,d_{m_{s1,j}}}\right\}$, constitutes the optimal strategy of LRRM 1. As for LRRM 2, since its problem is an integer programming, its optimal strategy is achieved by an exhaustive search as well.

Based on the above analysis, the algorithm presented in Algorithm 1 is introduced to find the SE of the formulated game. In Algorithm 1, the GRRM needs only some coarse information such as the number of UEs in slice $s2$ and the performance feedbacks from the two LRRMs to search for its optimal strategy, which avoids the collection of global CSI. Meanwhile, the number of its possible resource allocation strategies does not depend on the network scale of each slice. Nevertheless, the algorithm possesses high complexity, especially for both LRRMs. For example, the exhaustive search complexity under a given $\mathbf{d}$ for LRRM 1 is $\mathcal{O}\left(2^{|\mathcal{D}_{s1}|M_{s1}K}\right)$, which is related to the number of the allocated subchannels and the number of the UEs and RRHs of slice $s1$. Hence, if the network scale is large, it may be unrealistic for the LRRMs to get their optimal strategy within limited decision making time. Considering this fact, it is reasonable to regard the GRRM and LRRMs as players with bounded rationality that aim to find satisfactory solutions instead of optimal ones. We will develop low complexity algorithms based on this setting in Section V.

Algorithm 1 Algorithm based on exhaustive search to identify the existence of SE and find SE (if SE exists)

1: **Stage 1:**
The GRRM generates all possible $\mathbf{d}$ satisfying $(c1)$, $(c2)$ and $(c3)$ in (9), whose set is denoted by $\mathcal{D}'$, and initializes an empty set $\mathcal{Q}$.
If $\mathcal{D}'$ is empty, Algorithm 1 terminates and there is no SE solution. Otherwise, go to Stage 2.

2: **Stage 2:**
For each subchannel allocation strategy $\mathbf{d} \in \mathcal{D}'$
1) The GRRM distributes strategy $\mathbf{d}$ to both LRRMs, and each LRRM reacts optimally based on exhaustive search.
2) If the corresponding problem is infeasible for the LRRM, the LRRM feeds back a signalling indicating problem infeasibility to the GRRM. Otherwise, the LRRM feeds back the value of its optimal utility to the GRRM.
3) When the utility values of both LRRMs are received, the GRRM adds the current $\mathbf{d}$ to $\mathcal{Q}$ and records its achieved utility denoted by $U_0(\mathbf{d})$.
End For

3: **Stage 3:**
If $\mathcal{Q}$ is nonempty, $\mathbf{d}^* = \arg\min_{\mathbf{d}\in\mathcal{Q}} U_0(\mathbf{d})$, and $\mathbf{d}^*$, together with the corresponding optimal strategies of both LRRMs, constitutes the SE of the game. If there are multiple SE strategies for the GRRM, the SE strategy leading to the minimum download latency of slice $s2$ is selected as the system operation point.

V. Low Complexity Algorithm Design for Game Players

In this section, low complexity algorithms are developed for both the GRRM and LRRMs to help them achieve local optimal solutions when they behave as players with bounded rationality who intend to find satisfactory solutions within limited time for decision making.

A. Low Complexity Algorithm Design for LRRM 1

From problem (7), it can be seen that the main complexity lies in the optimization of $w_{m_{s1,j},k,d}$ representing RRH association and subchannel allocation. Note that subchannel allocation is actually equivalent to identifying the subchannel sharing relationship among UEs, which can be considered from a coalitional game perspective. Specifically, each subchannel can correspond to a coalition, and the UE joining a certain coalition occupies the corresponding subchannel. A low complexity and effective approach to form a stable coalitional structure is to make UEs traverse coalitions based on the transfer order [42], and the definition of the order

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TVT.2019.2896586, IEEE Transactions on Vehicular Technology

depends on the system optimization objective. Since LRRM 1 aims at minimizing system transmit power consumption, it is assumed that a UE transfers from a coalition to another coalition if and only if the overall power consumption strictly decreases.

When the subchannel allocation is fixed, the remaining task is to design a low complexity scheme to jointly optimize the UE-RRH association and precoding. To facilitate theoretical analysis, we define a UE-RRH association matrix $\mathbf{E}$ whose (j,k)-th element $e_{m_{s1,j},k}$ is a 0-1 variable, which equals to 1 if UE $m_{s1,j}$ associates with RRH k and equals to 0 otherwise. Then, the problem of LRRM 1 under fixed subchannel allocation is given by

$$\begin{aligned}
&\min_{\left\{\mathbf{v}_{m_{s1,j},k,d_{m_{s1,j}}}\right\}\left\{e_{m_{s1,j},k}\right\}} \sum_{m_{s1,j}\in\mathcal{M}_{s1}}\sum_{k}\left\|\mathbf{v}_{m_{s1,j},k,d_{m_{s1,j}}}\right\|_2^2\\
&(f1)\ \sqrt{\sum_{m_{s1,i},d_{m_{s1,i}}=d_{m_{s1,j}}}\left|\mathbf{h}_{m_{s1,j},d_{m_{s1,j}}}^H\mathbf{v}_{m_{s1,i}}\right|^2+\sigma^2}\\
&\le\sqrt{1+\frac{1}{\gamma_{m_{s1,j}}}}Re\left\{\mathbf{h}_{m_{s1,j},d_{m_{s1,j}}}^H\mathbf{v}_{m_{s1,j}}\right\},\forall m_{s1,j},\\
&(f2)\ \mathrm{Im}\left\{\mathbf{h}_{m_{s1,j},d_{m_{s1,j}}}^H\mathbf{v}_{m_{s1,j}}\right\}=0,\forall m_{s1,j},\\
&(f3)\ \sum_{m_{s1,j}}\left\|\mathbf{v}_{m_{s1,j},k,d_{m_{s1,j}}}\right\|_2^2\le p_{\max},\forall k,\\
&(f4)\ \sum_{m_{s1,j}}e_{m_{s1,j},k}\le o_k,\ \forall k,\\
&(f5)\ \mathbf{v}_{m_{s1,j},k,d_{m_{s1,j}}}=0,\ \ if\ \ e_{m_{s1,j},k}=0,\\
&(f6)\ e_{m_{s1,j},k}\in\{0,1\}.
\end{aligned}\tag{14}$$

To develop low complexity design for RRH association and precoding, we first analyze the characteristic of problem (14). Specifically, define the feasible set of precoding for problem (14) under association matrix $\mathbf{E}$ as $\mathcal{V}_{\mathbf{E}}$, and the corresponding optimal power consumption is denoted by $u_{\mathbf{E}}^*$. If an element $e_{m_{s1,j},k}=1$ in $\mathbf{E}$ is set to 0 and the new association matrix is denoted as $\mathbf{E}^-$, we have $\mathcal{V}_{\mathbf{E}^-}\subseteq\mathcal{V}_{\mathbf{E}}$ and hence $u_{\mathbf{E}}^*\le u_{\mathbf{E}^-}^*$. This means more active UE-RRH links can contribute to a better slice performance, which motivates us to introduce Algorithm 2 to have as many active links as possible meanwhile keeping the feasibility of constraint $(f4)$ of problem (14). The whole algorithm is mainly composed of two parts. The first part starts with activating all UE-RRH links and gradually drops off weak links to satisfy fronthaul capacity constraints including **Stage 1**-**Stage 4**, while the second part, including **Stage 5**-**Stage 6**, tries to reactivate links dropped off in the first part. The metric to identify weak links, which is called contribution index in the following, is given by

$$\beta_{m_{s1,j},k}=\frac{\left|\mathbf{h}_{m_{s1,j},k,d_{m_{s1,j}}}^H\mathbf{v}_{m_{s1,j},k,d_{m_{s1,j}}}\right|^2}{\left|\sum_{k'\in\mathcal{K}}\mathbf{h}_{m_{s1,j},k',d_{m_{s1,j}}}^H\mathbf{v}_{m_{s1,j},k',d_{m_{s1,j}}}\right|^2},\tag{15}$$

which shows the contribution of RRH k on the received signal strength of UE $m_{s1,j}$. It should be highlighted that problem (14) is reduced to problem (13) under fixed UE-RRH association. Therefore, under a fixed association matrix meeting fronthaul capacity constraints $(f4)$ in (14), checking the feasibility of problem (14) is equivalent to checking the feasibility of problem (13). In other words, problem (14) is feasible if and only if the association matrix satisfies fronthaul capacity constraints and meanwhile problem (13) is also feasible under this association matrix. Based on this finding, Stage 3 in Algorithm 2 is designed, which checks both the violation of fronthaul capacity constraints and the feasibility of problem (13).

Note that the number of possible UE-RRH links checked in Algorithm 2 is at most $2M_{s1}K$. Since each check may include a precoding optimization whose complexity is $\mathcal{O}\left(\left(KM_{s1}\right)^{3.5}N^{3.5}\right)$, the total complexity of Algorithm 2 is $\mathcal{O}\left(\left(KM_{s1}\right)^{4.5}N^{3.5}\right)$. After the algorithm terminates, the RRH association outcome is a local optimal solution to problem (14), and rigorous proof is as follows.

Theorem *2:* The final outcome achieved by Algorithm 2 is a local optimal solution to problem (14) in the sense that activating or inactivating any RRH-UE link can not further improve the slice performance while keeping all the constraints satisfied.

Proof: Denote the final association matrix output by Algorithm 2 as $\mathbf{E}$ which must be feasible to problem (14). As discussed above, inactivating a link in $\mathbf{E}$ can not contribute to a lower transmit power consumption due to a smaller feasible set of precoding for problem (13). Next, we prove that activating a link in $\mathbf{E}$ can not further improve slice performance as well. First, if $\mathbf{E}$ is a matrix with all elements equal to 1, it is obvious that no link can be activated, and hence $\mathbf{E}$ must be local optimal.

Then, we discuss the case where there exists at least one zero element in $\mathbf{E}$. Assume that there exists another association matrix $\mathbf{E}^+$ that is got based on $\mathbf{E}$ by activating a UE-RRH link $(m_{s1,j},k)$ that has been dropped off, and meanwhile meets the fronthaul capacity constraints for all RRHs. Define the set of active links in association matrix $\mathbf{E}$ as $\mathcal{S}_{active}^{\mathbf{E}}$, and then we have $\mathcal{S}_{active}^{\mathbf{E}}\subset\mathcal{S}_{active}^{\mathbf{E}^+}$. According to Algorithm 2, the UE-RRH link $(m_{s1,j},k)$ must be checked in **Stage 5**, and we denote the corresponding association matrix as $\mathbf{E}'$. From the algorithm, it can be known that $\mathbf{E}'$ does not satisfy fronthaul capacity constraints for all RRHs, because the UE-RRH link $(m_{s1,j},k)$ will be activated in $\mathbf{E}$ otherwise. But on the other hand, since $\mathcal{S}_{active}^{\mathbf{E}'}\subseteq\mathcal{S}_{active}^{\mathbf{E}}$ and $\mathcal{S}_{active}^{\mathbf{E}}\subset\mathcal{S}_{active}^{\mathbf{E}^+}$, we have $\mathcal{S}_{active}^{\mathbf{E}'}\subset\mathcal{S}_{active}^{\mathbf{E}^+}$, and hence $\mathcal{S}_{active}^{\mathbf{E}'}$ must satisfy all the fronthaul constraints as well, which makes a contradiction. Thus, activating any inactive link in $\mathbf{E}$ must lead to an association matrix violating fronthaul capacity constraints. Finally, we can conclude that the outcome achieved by Algorithm 2 must be local optimal. ■

Finally, an iterative algorithm based on coalitional game with transfer order is proposed for LRRM 1, which is described in Algorithm 3. Since the numbers of available subchannels and UEs are both limited, the possible subchannel allocation results are limited, and thus the number of possible coalitional structures is limited. Note that each successful transfer guarantees that the transmit power of slice $s1$ strictly decreases, which leads to a totally new coalitional structure. Therefore, the algorithm finally converges after a finite number of iterations. Moreover, according to the algorithm, its convergence indicates a stable solution to the subchannel

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TVT.2019.2896586, IEEE Transactions on Vehicular Technology

Algorithm 2 Resource allocation under pre-determined subchannel allocation for LRRM 1

1: **Stage 1:**
LRRM 1 initializes a UE-RRH association matrix $\mathbf{E}$ with all the elements equal to 1 and a set $\mathcal{S}_1=\{(m_{s1,j},k)\,|m_{s1,j}\in\mathcal{M}_{s1},k\in\mathcal{K}\}$. Then, solve problem (13) and check the fronthaul capacity constraint for each RRH under the association matrix $\mathbf{E}$.
- If problem (13) is infeasible, it can be concluded that no feasible RRH association exists for problem (14).
- If problem (13) is feasible and meanwhile the fronthaul capacity constraints are all met, then $\mathbf{E}$ together with the optimal precoding is output as the final solution.
- If problem (13) is feasible but fronthaul capacity constraints for certain RRHs are violated, LRRM 1 calculates (15) for each UE-RRH link, and go to **Stage 2**.

2: **Stage 2:**
Find $(m_{s1,j'},k')=\arg\min_{(m_{s1,j},k)\in\mathcal{S}_1}\beta_{m_{s1,j},k}$, and then set the (j',k')-th element in $\mathbf{E}$ to 0. In addition, let $\mathcal{S}_1\leftarrow\mathcal{S}_1/\{(m_{s1,j'},k')\}$.

3: **Stage 3:**
Solve problem (13) and check fronthaul capacity constraints.
If problem (13) is infeasible, reset the value of the (j',k')-th element in $\mathbf{E}$ to 1. Then go to **Stage 4**.
Else if problem (13) is feasible but the fronthaul capacity constraints are unsatisfied for some RRHs, update contribution index for each UE-RRH link in the set $\mathcal{S}_1$. Then go to **Stage 4**.
Else if problem (13) is feasible and meanwhile fronthaul capacity constraints hold for all RRHs, initialize a set $\mathcal{S}_2=\left\{(m_{s1,j},k)\,|e_{m_{s1,j},k}=0\right\}$ and go to **Stage 5**.

4: **Stage 4:**
If $\mathcal{S}_1=\phi$, the algorithm terminates and no feasible solution is obtained.
Else if go back to **Stage 2**.

5: **Stage 5:**
Select a UE-RRH link $(m_{s1,j'},k')$ in $\mathcal{S}_2$, and $\mathcal{S}_2\leftarrow\mathcal{S}_2/\{(m_{s1,j'},k')\}$. Set the value of the (j',k')-th element in $\mathbf{E}$ to 1.
If fronthaul capacity constraints do not hold for all RRHs with current $\mathbf{E}$, reset the value of the (j',k')-th element in $\mathbf{E}$ to 0.

6: **Stage 6:**
If $S_2=\phi$, the algorithm terminates.
Else if go back to **Stage 5**.

allocation in the sense that the system transmit power cannot be strictly decreased by changing any UE's subchannel allocation unilaterally. Considering Theorem 2, the resource allocation outcome resulting from Algorithm 3 can be claimed to be local optimal. Specifically, keeping RRH association and precoding fixed, changing any UE's subchannel allocation can not improve the slice performance, while activating or dropping off any UE-RRH link can not improve the slice performance as well when keeping the subchannel allocation fixed. Finally, the total complexity for the above algorithm in each iteration is $\mathcal{O}\left(|\mathcal{D}_{s1}|\,{M_{s1}}^{5.5}K^{4.5}N^{3.5}\right)$.

Algorithm 3 Resource allocation algorithm based on transfer order for LRRM 1

1: **Stage 1:**
LRRM 1 initializes a coalitional structure $\Pi_{ini}=\left\{\pi_1,\pi_2,...,\pi_{|\mathcal{D}_{s1}|}\right\}$ with $\pi_d\subseteq\mathcal{M}_{s1}$, $\pi_1\cup\pi_2\cdots\cup\pi_{|\mathcal{D}_{s1}|}=\mathcal{M}_{s1}$, and $\pi_d\cap\pi_{d'}=\phi,\ \forall d\neq d'$. Denote the index of the coalition to which UE $m_{s1,j}$ belongs as $d_{m_{s1,j}}\in\{1,2,...,|\mathcal{D}_{s1}|\}$. Compute $U_1(\Pi_{ini})$ by solving the RRH association and precoding optimization problem (14) using Algorithm 2.

2: **Stage 2:**
For $j=1:M_{s1}$
For $d=1:|\mathcal{D}_{s1}|$
If $d\neq d_{m_{s1,j}}$
$\Pi_{current}=\Pi_{ini}/\left\{\pi_d,\pi_{d_{m_{s1,j}}}\right\}\cup\left\{\pi_d\cup\{m_{s1,j}\},\pi_{d_{m_{s1,j}}}/\{m_{s1,j}\}\right\}$. Compute $U_1(\Pi_{current})$ using Algorithm 2.
If $U_1(\Pi_{ini})>U_1(\Pi_{current})$ (transfer condition)
$\Pi_{ini}\leftarrow\Pi_{current}$ and $d_{m_{s1,j}}\leftarrow d$.
End If
End If
End For
End For

3: **Stage 3:**
Repeat **Stage 2** until the coalitional structure Π_{ini} converges.

B. Low Complexity Algorithm Design for LRRM 2

The optimization problem (8) of LRRM 2 is actually a matching problem between UEs and resources, which motivates the adoption of matching theory to develop a low complexity algorithm. According to the constraints in (8), each UE can be allocated with only one subchannel-FAP pair and each subchannel-FAP pair can serve only one UE. Hence, a one-to-one matching problem between UEs and subchannel-FAP pairs can be formulated. Define the set of all the subchannel-FAP pairs as $\mathcal{A}$ with $|\mathcal{A}|=|\mathcal{D}_{s2}|\,L$. Then, the concerned matching problem is formally defined as follows.

Definition *2:* A matching μ is a one-to-one mapping between UEs and subchannel-FAP pairs satisfying
1.If $\mu(m_{s2,j})=a$, then $\mu(a)=m_{s2,j}$, $\forall m_{s2,j}\in\mathcal{M}_{s2}$ and $\forall a\in\mathcal{A}$, and vice versa.
2.$|\mu(m_{s2,j})|=1,\forall m_{s2,j}\in\mathcal{M}_{s2}$.
3.$|\mu(a)|\leq 1,\forall a\in\mathcal{A}$.

The first condition states that a matching establishes a mutual relationship between the elements in two different sets. The second condition requires that each UE must be mapped to a subchannel-FAP pair, otherwise the average download latency will be infinite. The third condition guarantees that each subchannel-FAP pair can serve at most one UE. Moreover, the

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TVT.2019.2896586, IEEE Transactions on Vehicular Technology

download latency of each UE calculated by (6) depends on not only its own matching to subchannel-FAP pairs but also the matching between other UEs and subchannel-FAP pairs, which causes externality of the formulated matching problem. Hence, the classic deferred acceptance algorithm cannot be applied [45]. Meanwhile, there is no guarantee on the existence of a pair-wise stable matching [46]. Therefore, the concept of swap matching is incorporated to effectively tackle the matching problem, whose definition is given as follows.

Definition *3:* Assume that under matching μ, $\mu\left(m_{s2,j}\right)=a$ and $\mu\left(m_{s2,j'}\right)=a'$. Then, $\mu^{a,a'}_{m_{s2,j},m_{s2,j'}}=\mu/\left\{\left(m_{s2,j},a\right),\left(m_{s2,j'},a'\right)\right\}\cup\left\{\left(m_{s2,j},a'\right),\left(m_{s2,j'},a\right)\right\}$ is a swap matching.

It should be highlighted that one of the UEs involved in the swap can be a dummy UE, which means that no UE is served by the subchannel-FAP pair it matches with, and a UE can be mapped to this pair directly. However, considering that an FAP can serve only finite number of UEs, there is a special case for the definition of swap matching. Specifically, suppose that a UE $m_{s2,j}$ with $\mu\left(m_{s2,j}\right)=a$ intends to swap to a subchannel-FAP pair a' not occupied by any UE, and the FAP in pair a' does not associate with UE $m_{s2,j}$ currently. If the quota of the FAP has been full, i.e., the constraint $(b3)$ in problem (8) takes equality, to keep the feasibility of the resource allocation solution, a UE $m_{s2,j''}$ served by the FAP should be switched to the subchannel-FAP pair a when UE $m_{s2,j}$ swaps to subchannel-FAP pair a'.

Since the LRRM aims to minimize the average download latency of UEs, the swap condition is defined as follows.

Definition *4:* (Swap Condition) Denote the subchannels included in subchannel-FAP pair a and a' by d_a and $d_{a'}$, respectively, and define $\mathcal{D}_{swap}=\{d_a\}\cup\{d_{a'}\}$. Then, a swap matching $\mu^{a,a'}_{m_{s2,j},m_{s2,j'}}$ is preferred by LRRM 2 to the current matching μ if

$$\begin{aligned}&\sum_{m_{s2,j},d_{\mu\left(m_{s2,j}\right)}\in\mathcal{D}_{swap}} t_{m_{s2,j}}\left(\mu^{a,a'}_{m_{s2,j},m_{s2,j'}}\right)\\ &<\sum_{m_{s2,j},d_{\mu\left(m_{s2,j}\right)}\in\mathcal{D}_{swap}} t_{m_{s2,j}}\left(\mu\right).\end{aligned}\tag{16}$$

Note that the download latency for UE $m_{s2,j}$ is expressed as a function of the current matching due to the existence of externality. For the special case of swap matching discussed before, denote the subchannel allocated to UE $m_{s2,j''}$ by $d_{a''}$, and then the set of subchannels related to the swap is now given by $\mathcal{D}_{swap}=\{d_a\}\cup\{d_{a'}\}\cup\{d_{a''}\}$.

Based on the above definitions, a low complexity resource allocation algorithm for LRRM 2 is presented in Algorithm 4. From the swap condition, it can be seen that each swap operation leads to a strict decrease of the system average download latency, and hence a completely new matching will be generated after each swap. Because of the limited number of all possible UE-resource matchings, the algorithm will converge to a final matching after finite iterations. Moreover, the convergence of the matching means a local optimal resource allocation outcome in the sense that any swap will not improve the latency performance of slice $s2$. The complexity of the algorithm for each iteration is $\mathcal{O}\left(M_{s2}\left|\mathcal{D}_{s2}\right|Lq\right)$, where q is the maximal number of UEs that can be served by an FAP.

Algorithm 4 Resource allocation algorithm based on swap matching for LRRM 2

1: **Stage 1:**
LRRM 2 generates the set of all possible subchannel-FAP pairs $\mathcal{A}$ and initializes a feasible UE-resource matching μ satisfying constraints in problem (8). Denote the i-th subchannel-FAP pair in set $\mathcal{A}$ as a_i.

2: **Stage 2:**
For $j=1:M_{s2}$
For $i=1:|\mathcal{A}|$
If $\mu\left(m_{s2,j}\right)\neq a_i$
Check whether $\mu^{a,a_i}_{m_{s2,j},\mu(a_i)}$ is preferred according to the swap condition. Especially, when UE $m_{s2,j}$ is not associated with the FAP included in subchannel-FAP pair a_i not matched to any UE, and meanwhile the quota of the FAP has been achieved, one of the q UEs served by the FAP should be swaped to the subchannel-FAP pair currently matched to UE $m_{s2,j}$ to keep a feasible matching.
If Swap condition holds
$\mu\leftarrow\mu^{a,a_i}_{m_{s2,j},\mu(a_i)}$.
End If
End If
End For
End For

3: **Stage 3:**
Repeat **Stage 2** until the UE-resource matching μ converges.

C. Low Complexity Algorithm Design for the GRRM

In this subsection, a low complexity algorithm will be developed for the GRRM to avoid the exhaustive complexity, and hence good scalability can be achieved when the number of subchannels is very large. Specifically, a concept similar to swap matching adopted in the above subsection is first formally defined as follows.

Definition *5:* (Resource Swap Operation) Define the first kind of swap operations of the GRRM as $\mathcal{D}_{s1}/\left\{d_{s1}\right\}\cup\left\{d_{s2}\right\}$ and $\mathcal{D}_{s2}/\left\{d_{s2}\right\}\cup\left\{d_{s1}\right\}$, where $d_{s1}\in\mathcal{D}_{s1}$ and $d_{s2}\in\mathcal{D}_{s2}$, and define the second kind of swap operations of the GRRM as $\mathcal{D}_{s1}\cup\left\{d_{s2}\right\}$ with $\mathcal{D}_{s2}/\left\{d_{s2}\right\}$ or $\mathcal{D}_{s2}\cup\left\{d_{s1}\right\}$ with $\mathcal{D}_{s1}/\left\{d_{s1}\right\}$.

Then, the following condition is given to identify whether to execute swap operations.

Definition *6:* (Swap Operation Condition) A swap operation is performed if and only if the utility of the GRRM U_0 is strictly decreased.

Based on these definitions, our proposed low complexity resource allocation scheme for the GRRM is elaborated in Algorithm 5. According to the swap operation condition, each update of subchannel allocation between slices means a strict improvement on the utility of the GRRM U_0. Since radio resources are limited, there is a lower bound of U_0. Hence, after finite subchannel allocation updates, Algorithm 5 will

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TVT.2019.2896586, IEEE Transactions on Vehicular Technology

Algorithm 5 Resource allocation algorithm based on swap operations for the GRRM

1: **Stage 1:**
Initially, the GRRM allocates one subchannel to slice $s2$, while the remaining subchannels are allocated to slice $s1$. After being informed about the resource allocation result, LRRM 1 and LRRM 2 perform local resource optimization, and then report slice performances to the GRRM, based on which the GRRM identifies its initial utility U_0 according to (9).

2: **Stage 2:**
For subchannel $d = 1 : D$:
For slice $i = s1 : s2$:
For subchannel $d_i \in \mathcal{D}_i$:
The GRRM tries the first kind of swap operations involving subchannel d and d_i, and notifies the LRRMs of slices about the subchannel allocation.
LRRM 1 and LRRM 2 do resource optimization, and then feed back slice performances to the GRRM.
If the swap operation condition holds:
Update subchannel allocation between slices and go back to the second **For** loop.
End If
End for
If no update of subchannel allocation between slices occurs:
Try the second kind of swap operations, and notify the LRRMs of slices about the subchannel allocation.
LRRM 1 and LRRM 2 do resource optimization, and then feed back slice performances to the GRRM.
If the swap operation condition holds:
Update subchannel allocation between slices and go back to the second **For** loop.
End If
End If
End For
End For

3: **Stage 3:**
Repeat **Stage 2** until the subchannel allocation between slices converges.

converge to a final resource allocation strategy. In addition, the strategy is local optimal in the sense that no swap operations can further improve the utility of the GRRM. The complexity of the algorithm in each iteration is $\mathcal{O}\left(D^2\right)$, showing a good scalability when the number of subchannels is large. Note that Algorithm 5 can be adopted to allocate any type of resources between two slices with any performance metrics. For example, each subchannel in this algorithm can be replaced by a resource bundle including radio resources, caching resources and computing resources. Moreover, once the GRRM reaches its local optimal strategy by Algorithm 5, it has no incentive to change the strategy according to the assumption of bounded rationality. Then, after each LRRM plays its local optimal strategy given the strategy of the GRRM, a weak version of SE is reached following the definition of SE.

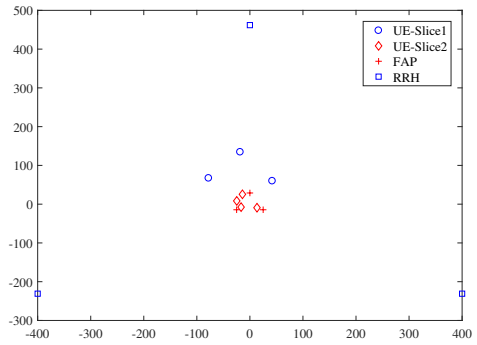

Fig. 2. Simulation scenario.

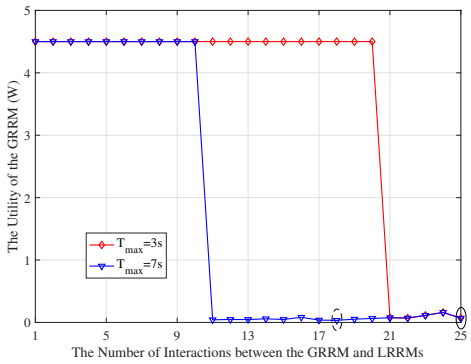

Fig. 3. The utility evolution of the GRRM.

VI. Simulation Results And Analysis

In all the simulations, the bandwidth of each subchannel is 180kHz, and the channel coefficient of each link is composed of path loss and fast fading. The pathloss model is d^{-2} with d being the distance between the two nodes in the same link. The fast fading is modeled as an independent complex Gaussian random variable distributed as $\mathcal{CN}(0, 1)$. The maximal transmit power of each RRH is 1.5 W, and the number of antennas in each RRH is 2. The noise power is set to 10^{-13} W, and the SINR requirement of UEs in slice $s1$ is taken as 5 dB. The transmit power of each FAP over each subchannel is 125 mW, and each FAP can serve at most 5 UEs. In addition, it is assumed that the files requested by UEs in slice $s2$ are available at the caches of all FAPs with the size of each file set to 10Mbits. Considering the randomness of user content requests, the assumption for 100% cache hit rate under a specific realization of user content requests is reasonable. However, note that our proposed resource allocation algorithms for slice $s2$ can fit into any cache hit situations.

A. Achieving SE via Exhaustive Search

The simulation scenario in this part consists of two F-RAN slices shown in Fig. 2, where the axis unit is in meters. Meanwhile, we suggest that five subchannels are to be allocated by the GRRM, and each fronthaul link can support at most two UEs. Note that the reason for adopting such a small scale topology in Fig. 2 is to facilitate the calculation of SE by exhaustive search.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TVT.2019.2896586, IEEE Transactions on Vehicular Technology

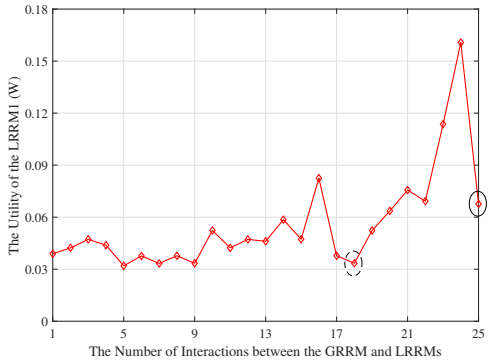

Fig. 4. The utility evolution of LRRM 1.

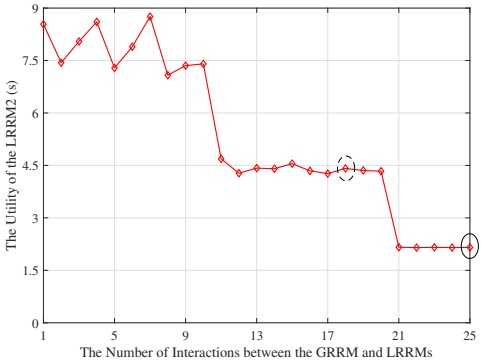

Fig. 5. The utility evolution of LRRM 2.

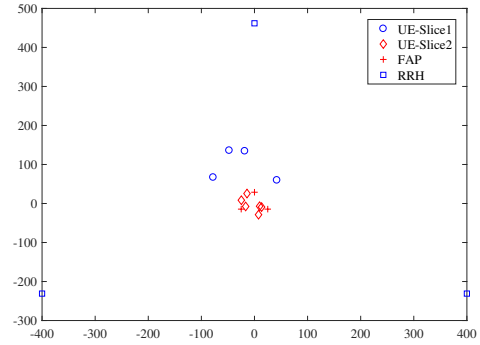

Fig. 6. Simulation scenario.

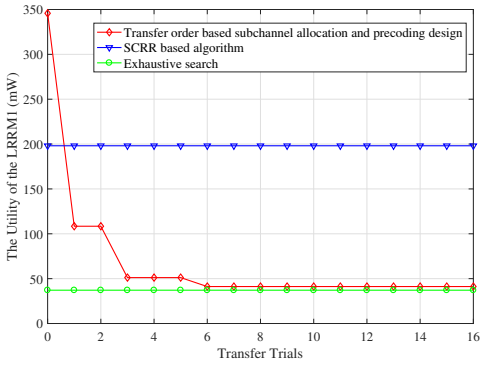

Fig. 7. The performance of the transfer order based resource allocation algorithm for LRRM 1.

Figs. 3-5 show the evolution of the utilities of the GRRM, LRRM 1 and LRRM 2 during the GRRM searches for its best strategy, and the total number of interactions between the GRRM and LRRMs is 25, since the number of subchannel allocation strategies of the GRRM satisfying the constraints $(c1)$-$(c3)$ in (9) is 25. Each interaction includes the broadcast of the GRRM's strategy to LRRMs, the resource optimization of LRRMs under that given strategy, their performance feedback to the GRRM, and the decision of the GRRM on whether to record the current strategy or not. Moreover, if a subchannel allocation strategy becomes infeasible to any player's optimization problem, the utility of the GRRM is set to 4.5. From Figs. 3-5, it can be seen that more subchannel allocation strategies become infeasible when the latency requirement of slice $s2$ becomes more stringent. Meanwhile, the SE points under $T_{max} = 3s$ and $T_{max} = 7s$ are marked by ellipses with solid line and dashed line, respectively, and it can be found from Fig. 4 that more transmission power is consumed by the slice $s1$ when $T_{max} = 3s$, which demonstrates a performance tradeoff between the two slices. Specifically, this is because the GRRM has to allocate more subchannels to the LRRM 2 to satisfy its more stringent performance requirement. Hence, there are fewer subchannels available for LRRM 1, leading to more severe multi-user interference.

B. The Effectiveness of The Low Complexity Algorithms for LRRMs

In this subsection, we compare the performance of the low complexity resource allocation schemes developed for LRRMs with exhaustive search and other low complexity baselines. Specifically, two heuristic resource allocation schemes are adopted in our simulation study:

- ***Static clustering and round-robin based subchannel allocation (SCRR)***: This scheme is adopted as the baseline for Algorithm 3. Specifically, following the Algorithm 3 in [44], the RRH cluster for each UE is first formed, and then each UE is allocated with a subchannel in a round-robin like fashion. After RRH association and subchannel allocation are determined, a precoding design problem with the same form of problem (13) will be solved with CVX.
- ***Deferred acceptance based node association and round-robin based subchannel allocation (DARR)***: This scheme is adopted as the baseline for Algorithm 4. Specifically, UEs associate with FAPs in slice $s2$ based on received signal strength, and then each node allocates each UE with a subchannel leading to the highest SINR for this UE in a round-robin like manner.

Moreover, the simulation scenario is enlarged on the basis of that in Fig. 2, where slice $s1$ has 4 UEs while slice $s2$ has 6 UEs.

Fig. 7 evaluates the performance of the proposed transfer order based resource allocation algorithm for LRRM 1, i.e., Algorithm 3, where the number of subchannels is set to 3, and the capacity of each fronthaul link is set to 3. From Fig. 7, it can seen that the transmit power consumption of slice $s1$ after each trial is not increased, which is guaranteed by the transfer

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TVT.2019.2896586, IEEE Transactions on Vehicular Technology

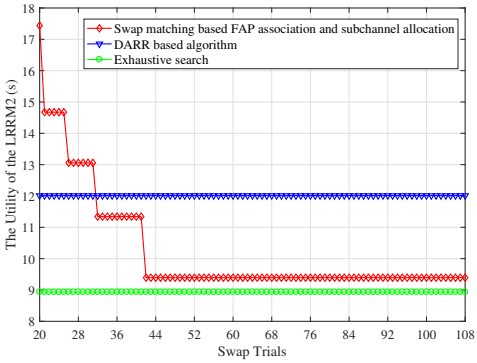

Fig. 8. The performance of the swap matching based resource allocation algorithm for LRRM 2.

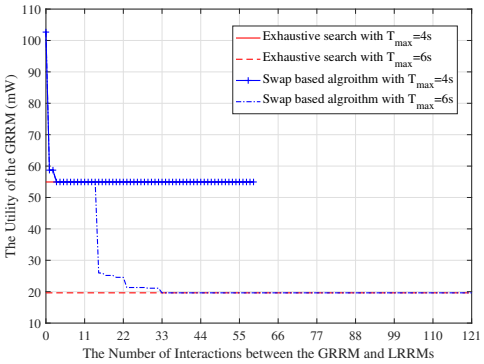

Fig. 9. The performance of the swap operations based resource allocation algorithm for the GRRM.

order. In addition, a huge gap between the performance of the proposed scheme and that of SCRR is observed, and this is because the static clustering and round-robin like subchannel allocation in SCRR do not well handle the multi-user interference. At last, the high complexity exhaustive search method can only decrease the transmit power consumption by no more than 10%, showing the competitive performance of our proposal.

Fig. 8 illustrates the performance of the proposed swap matching based resource allocation algorithm for LRRM 2, i.e., Algorithm 4, where the number of subchannels is set to 3. First, it can be found that the performance of slice $s2$ is continuously improved with swap operations going on. Moreover, it is observed that the algorithm can deliver a near optimal solution that leads to just about 5% average download latency increase compared with the optimal solution. Meanwhile, the improvement of our proposed scheme relative to the performance of DARR is significant.

C. The Effectiveness of The Low Complexity Algorithm for the GRRM

To verify the effectiveness of the low complexity proposal for the GRRM, a larger scale network on the basis of Fig. 6 is considered, where there are 13 UEs in slice $s2$ with each FAP capable of serving at most 5 UEs, and the GRRM needs to decide the allocation of 11 subchannels. The whole resource allocation procedure follows Algorithm 5, where each LRRM

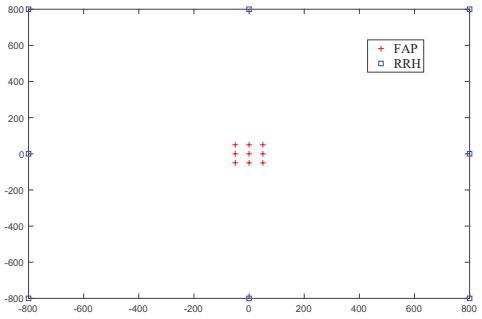

Fig. 10. Large scale simulation scenario.

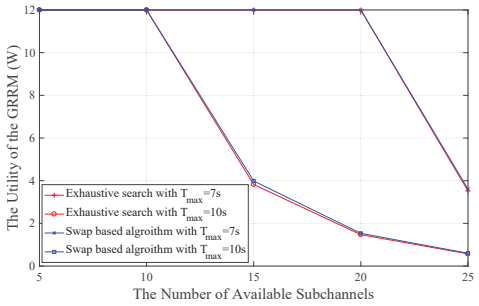

Fig. 11. The utility of the GRRM with the increase of available subchannels.

acts as a player with bounded rationality who derives a local optimal strategy using the developed low complexity resource allocation algorithm given the strategy of the GRRM. From Fig. 9, we can see that the proposal can achieve near optimal performance under different latency performance requirements of slice $s2$ but with much lower computation complexity than exhaustive search. In addition, a performance tradeoff between slices is observed similar to that in Subsection A.

To very the optimality and check the convergence time of Algorithm 5 under a large scale setting, the scenario in Fig. 10 is considered, where slice 1 has 8 RRHs while slice 2 has 9 FAPs. Meanwhile, there are 25 UEs in slice $s1$ uniformly distributed in a circle with the radius of 200m, whose center is the origin, and there are also 25 UEs in slice $s2$ uniformly distributed within a circle with the radius being 100m and taking the origin as the center. To accommodate the large number of UEs, fronthaul capacity of each RRH is set to 8.

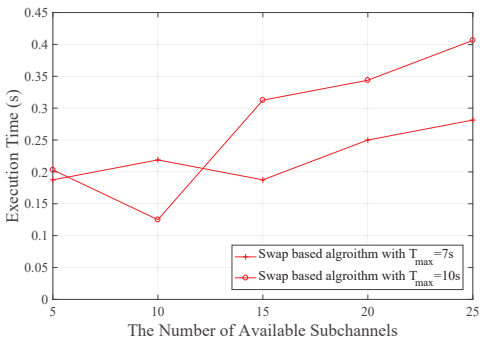

Fig. 12. The execution time of Algorithm 5 with the increase of available subchannels.

Fig. 11 compares the performance of Algorithm 5 with exhaustive search under different number of subchannels, and note that the utility of the GRRM is set to the maximal total transmit power of RRHs, which is 12 here, if the latency performance of slice $s2$ is not met. First, it can seen that our proposed swap based inter-slice resource allocation algorithm can reach near optimal performance. Second, when the number of subchannels increases, the utility of the GRRM is improved. This is because more subchannels provide more opportunity to satisfy the latency requirement of slice $s2$, and meanwhile more subchannels can be allocated to slice $s1$, which can reduce interference and thus enables slice $s1$ to deliver high data rate with lower transmit power consumption. Third, the utility of the GRRM is also improved when the latency requirement of slice $s2$ is less stringent. This is because less stringent latency requirement of slice $s2$ makes more subchannels available to slice $s1$. Moreover, the execution time of Algorithm 5 is demonstrated in Fig. 12, and the time has excluded the decision-making time of LRRMs. When the number of subchannels exceeds 15, more subchannels lead to longer execution time due to more possible inter-slice subchannel allocation strategies. Moreover, when the number of subchannels exceeds 15, less stringent performance requirement of slice $s2$ causes a larger execution time under fixed number of subchannels, since more inter-slice allocation strategies, which can meet the latency requirement of slice $s2$, exist.

VII. Conclusions

In this paper, radio resource allocation between two network slices with heterogenous performance metrics in fog radio access networks has been investigated. Under a hierarchical radio resource allocation architecture, the resource allocation problem has been modeled as a Stackelberg game, where the GRRM with a strong position acts as the leader and the LRRMs of slices act as followers. In addition, the game has been proved to possess at least one Stackelberg equilibrium (SE) under certain conditions. Since the problems of followers are NP-hard and the strategy of the leader is discrete, deriving SE solution is very challenging. To deal with this problem, a method based on exhaustive search has been proposed to help the GRRM and LRRMs achieve the SE interactively. Moreover, inspired by game-theoretic algorithms, low complexity resource allocation schemes have been developed for both the GRRM and LRRMs when they behave as players with bounded rationality. Furthermore, the optimality and complexity of our proposals are all rigorously analyzed. Simulation study has demonstrated that there is a tradeoff between the performance of slices, and the low complexity algorithms can achieve highly competitive solutions.

At last, it should be noted that our proposals for the GRRM can be used to allocate any resource bundles between slices by substituting subchannels with resource bundles in Algorithm 1 and Algorithm 5. In the future, it is interesting to study hierarchical resource allocation for more than two slices with heterogenous performance metrics. Meanwhile, when LRRMs cheat the GRRM in terms of their achieved performance, the GRRM may not find the real optimal or local optimal strategy, which means SE may not be achieved. Faced with this issue, it is essential to explore incentive mechanism design to make LRRMs report their performance truthfully.

References

[1] M. Peng, S. Yan, K. Zhang, and C. Wang, "Fog computing based radio access networks: Issues and challenges," *IEEE Netw.*, vol. 30, no. 4, pp. 46-53, Jul. 2016.

[2] S. Yan, M. Peng, and W. Wang, "User access mode selection in fog computing based radio access networks," in *Proceedings of ICC*, Kuala Lumpur, Malaysia, May 2016, pp. 1-6.

[3] H. Xiang, M. Peng, Y. Cheng, and H. Chen, "Joint mode selection and resource allocation for downlink fog radio access networks supported D2D," in *Proceedings of QSHINE*, Taipei, China, Aug. 2015, pp. 177-182.

[4] J. Kang, O. Simeone, J. Kang, and S. Shamai, "Joint optimization of cloud and edge processing for fog radio access networks," *IEEE Wireless Commun.*, vol. 15, no. 11, pp. 7621-7632, Nov. 2016.

[5] C. Liang and F. R. Yu, "Wireless network virtualization: A survey, some research issues and challenges," *IEEE Commun. Surveys & Tutorials*, vol. 17, no. 1, pp. 358-380, FIRST QUARTER 2015.

[6] 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 & Tutorials*, vol. 18, no. 3, pp. 2282-2308, THIRD QUARTER 2016.

[7] M. Richart, J. Baliosian, J. Serrat, and J. Gorricho, "Resource slicing in virtual wireless networks: A survey," *Trans. Netw. and Service Mana.*, vol. 13, no. 3, pp. 462-476, Sep. 2016.

[8] K. Zhu and E. Hossain, "Virtualization of 5G cellular networks as a hierarchical combinatorial auction," *IEEE Trans. Mobile Comput.*, vol. 15, no. 10, pp. 2640-2654, Dec. 2015.

[9] K. Wang, H. Li, F. R. Yu, and W. Wei, "Virtual resource allocation in software-defined information-centric cellular networks with device-to-device communications and imperfect CSI," *IEEE Trans. Veh. Technol.*, vol. 65, no. 12, pp. 10011-10021, Dec. 2016.

[10] T. LeAnh, N. H. Tran, D. T. Ngo, and C. S. Hong, "Resource allocation for virtualized wireless networks with backhaul constraints," *IEEE Commun. Lett.*, vol. 21, no. 1, pp. 148-151, Jan. 2017.

[11] V. Jumba, S. Parsaeefard, M. Derakhshani, and T. Le-Ngoc, "Resource provisioning in wireless virtualized networks via massive-MIMO," *IEEE Wireless Commun. Lett.*, vol. 3, no. 4, pp. 237-240, Jun. 2015.

[12] V. Jumba, S. Parsaeefard, M. Derakhshani, and T. Le-Ngoc, "Dynamic resource provisioning with stable queue control for wireless virtualized networks," in *Proceedings of PIMRC*, Hongkong, China, Sep. 2015, pp. 1856-1860.

[13] L. Gao, P. Li, Z. Pan, N. Liu, and X. You, "Virtualization framework and VCG based resource block allocation scheme for LTE virtualization," in *Proceedings of VTC*, Nanjing, China, May 2016, pp. 1-6.

[14] S. Parsaeefard, R. Dawadi, M. Derakhshani, and T. Le-Ngoc, "Joint user-association and resource-allocation in virtualized wireless networks," *IEEE Access*, vol. 4, pp. 2738-2750, Apr. 2016.

[15] M. Richart, J. Baliosian, J. Serrat, and J. Gorricho, "Resource slicing in virtual wireless networks: A survey," *IEEE Trans. Netw. & Serv. Manag.*, vol. 13, no. 3, pp. 462-476, Sep. 2016.

[16] B. Ma, M. H. Cheung, V. W. S. Wong, and J. Huang, "Hybrid overlay/underlay cognitive femtocell networks: A game theoretic approach," *IEEE Trans. Wireless Commun.*, vol. 14, no. 6, pp. 3259-3270, Jun. 2015.

[17] F. Pantisano, M. Bennis, W. Saad, M. Debbah, and M. Latva-aho, "Interference alignment for cooperative femtocell networks: A game-theoretic approach," *IEEE Trans. Mobile Comput.*, vol. 12, no. 11, pp. 2233-2246, Nov. 2013.

[18] M. Ahmed, M. Peng, M. Abana, S. Yan, and C. Wang, "Interference coordination in heterogeneous small-cell networks: A coalition formation game approach," *IEEE Syst. J.*, vol. 12, no. 1, pp. 604-615, Mar. 2018.

[19] Y. Chen, B. Ai, Y. Niu, K. Guan, and Z. Han, "Resource allocation for device-to-device communications underlaying heterogeneous cellular networks using coalitional games," *IEEE Trans. Wireless Commun.*, vol. 17, no. 6, pp. 4163-4176, Jun. 2018.

[20] X. Kang, R. Zhang, and M. Motani, "Price-based resource allocation for spectrum-sharing femtocell networks: A stackelberg game approach," *IEEE J. Sel. Areas Commun.*, vol. 30, no. 3, pp. 538-549, Apr. 2012.

[21] P. Yuan, Y. Xiao, G. Bi, and Liren Zhang, "Toward cooperation by carrier aggregation in heterogeneous networks: A hierarchical game approach," *IEEE Trans. Veh. Technol.*, vol. 66, no. 2, pp. 1670-1683, Feb. 2017.

[22] S. Bu, F. R. Yu, and H. Yanikomeroglu, "Interference-aware energy-efficient resource allocation for OFDMA-based heterogeneous networks with incomplete channel state information," *IEEE Trans. Veh. Technol.*, vol. 64, no. 3, pp. 1036-1050, Mar. 2015.

[23] Y. Wang, X. Wang, and L. Wang, "Low-complexity stackelberg game approach for energy-efficient resource allocation in heterogeneous networks," *IEEE Commun. Lett.*, vol. 18, no. 11, pp. 2011-2014, Nov. 2014.

[24] S. Guruacharya, D. Niyato, D. I. Kim, and E. Hossain, "Hierarchical competition for downlink power allocation in OFDMA femtocell networks," *IEEE Trans. Wireless Commun.*, vol. 12, no. 4, pp. 1543-1553, Apr. 2013.

[25] L. Liang, G. Feng, and Y. Jia, "Game-theoretic hierarchical resource allocation for heterogeneous relay networks," *IEEE Trans. Veh. Technol.*, vol. 64, no. 4, pp. 1480-1492, Apr. 2015.

[26] R. Xie, F. R. Yu, H. Ji, and Y. Li, "Energy-efficient resource allocation for heterogeneous cognitive radio networks with femtocells," *IEEE Trans. Wireless Commun.*, vol. 11, no. 11, pp. 3910-3920, Nov. 2012.

[27] H. Dai, Y. Huang, J. Wang, and L. Yang, "Resource optimization in heterogeneous cloud radio access networks," *IEEE Commun. Lett.*, vol. 22, no. 3, pp. 494-497, Mar. 2018.

[28] M. Peng, Y. Wang, T. Dang, and Z. Yan, "Cost-efficient resource allocation in cloud radio access networks with heterogeneous fronthaul expenditures," *IEEE Trans. Wireless Commun.*, vol. 16, no. 7, pp. 4626-4638, Jul. 2017.

[29] M. Peng *et al.*, "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.

[30] Y. Sun, M. Peng, and S. Mao, "Deep reinforcement learning based mode selection and resource management for green fog radio access networks," *IEEE Internet Things J.*, Sep. 2018, doi: 10.1109/JIOT.2018.2871020, submitted for publication.

[31] J. Tang, R. Wen, T. Q. S. Quek, and M. Peng, "Fully exploiting cloud computing to achieve a green and flexible C-RAN," *IEEE Commun. Mag.*, vol. 55, no. 11, pp. 40-46, Nov. 2017.

[32] H. Xiang, W. Zhou, M. Daneshmand, and M. Peng, "Network slicing in fog radio access networks: Issues and challenges," *IEEE Commun. Mag.*, vol. 55, no. 12, pp. 110-116, Dec. 2017.

[33] Y. Sun, M. Peng, and H. Vincent Poor, "A distributed approach to improving spectral efficiency in uplink device-to-device enabled cloud radio access networks," *IEEE Trans. Commun.*, vol. 66, no. 12, pp. 6511-6526, Dec. 2018.

[34] M. Bennis, S. M. Perlaza, P. Blasco, Z. Han, and H. Vincent Poor, "Self-organization in small cell networks: A reinforcement learning approach," *IEEE Trans. Wireless Commun.*, vol. 12, no. 7, pp. 3202-3212, Jul. 2013.

[35] K. Senel and M. Akar, "A power allocation algorithm for multi-tier cellular networks with heterogeneous QoS and imperfect channel considerations," *IEEE Trans. Wireless Commun.*, vol. 16, no. 11, pp. 7184-7194, Nov. 2017.

[36] B. Yuksekkaya and C. Toker, "Power and interference regulated water-filling for multi-tier multi-carrier interference aware uplink," *IEEE Wireless Commun. Lett.*, vol. 7, no. 4, pp. 494-497, Aug. 2018.

[37] Y. Lin, W. Bao, W. Yu, and B. Liang, "Optimizing user association and spectrum allocation in HetNets: A utility perspective," *IEEE J. Sel. Areas Commun.*, vol. 33, no. 6, pp. 1025-1039, Jun. 2015.

[38] S. A. R. Naqvi *et al.*, "Energy-aware radio resource management in D2D-enabled multi-tier HetNets," *IEEE Access*, vol. 6, pp. 16610-16622, Mar. 2018.

[39] S. Ali, A. Ahmad, R. Iqbal, S. Saleem, and T. Umer, "Joint RRH-association, sub-channel assignment and power allocation in multi-tier 5G C-RANs," *IEEE Access*, vol. 6, pp. 34393-34402, Jun. 2018.

[40] C. Xu, M. Sheng, V. S. Varma, T. Q. S. Quek, and J. Li, "Wireless service provider selection and bandwidth resource allocation in multi-tier HCNs," *IEEE Trans. Commun.*, vol. 64, no. 12, pp. 5108-5124, Dec. 2016.

[41] Z. Han *et al.*, "Game theory in wireless and communication networks," Cambridge University Press, 2012.

[42] W. Saad, Z. Han, R. Zheng, M. Debbah, and H. V. Poor, "A college admissions game for uplink user association in wireless small cell networks," in *Proceedings of INFOCOM*, Toronto, ON, Canada, Apr. 2014, pp. 1096-1104.

[43] E. J. Candes, M. B. Wakin, and S. P. Boyd, "Enhancing sparsity by reweighted l-1 minimization," *Journal of Fourier analysis and applications*, vol. 14, no. 5-6, pp. 877-905, 2008.

[44] B. Dai and W. Yu, "Sparse beamforming and user-centric clustering for downlink cloud radio access network," *IEEE Access*, vol. 2, pp. 1326-1339, Nov. 2014.

[45] F. Pantisano, M. Bennis, W. Saad, S. Valentin, and M. Debbah, "Matching with externalities for context-aware user-cell association in small cell networks," in *Proceedings of Globecom Workshops*, Atlanta, GA, USA, Dec. 2013, pp. 4483-4488.

[46] A. E. Roth and M. A. O. Sotomayor, "Two-sided matching: A study in game-theoretic modeling and analysis," Cambridge University Press, Jun. 1992.

Yaohua Sun received the bachelor's degree (with first class Hons.) in telecommunications engineering (with management) from Beijing University of Posts and Telecommunications (BUPT), Beijing, China, in 2014. He is a Ph.D. student at the Key Laboratory of Universal Wireless Communications (Ministry of Education), BUPT. He has been reviewers for *IEEE Transactions on Communications*, *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*.

Mugen Peng (M'05, SM'11) received the Ph.D. degree in communication and information systems from the Beijing University of Posts and Telecommunications (BUPT), Beijing, China, in 2005. Afterward, he joined BUPT, where he has been a Full Professor since 2012. He has authored and coauthored over 100 refereed IEEE journal papers and over 300 conference proceeding papers. His main research areas include wireless communication theory, radio signal processing, cooperative communication, self-organization networking, heterogeneous networking, cloud communication, and Internet of Things. Dr. Peng was a recipient of the 2018 Heinrich Hertz Prize Paper Award, the 2014 IEEE ComSoc AP Outstanding Young Researcher Award, and the Best Paper Award in the JCN 2016, IEEE WCNC 2015, etc. He is currently or have been on the Editorial/Associate Editorial Board of the *IEEE Communications Magazine*, *IEEE ACCESS* and *IEEE Internet of Things Journal*.