Joint Resource Allocation and Caching Placement for Network Slicing in Fog Radio Access Networks

Liya Tang, Xian Zhang, Hongyu Xiang, Yaohua Sun, and Mugen Peng
Key Laboratory of Universal Wireless Communication, Ministry of Education,
Beijing University of Posts and Telecommunications, Beijing, 100876, China.
Email: tangliya@bupt.edu.cn, pmg@bupt.edu.cn

***Abstract*—To satisfy diverse use cases and business models in fifth generation (5G) wireless communication, network slicing in fog radio access networks (F-RANs) is proposed, which provides a cost efficient networking in a convenient way. However, the resource management for network slices is challenging, especially when the edge caching is utilized to alleviate the fronthaul burden and reduce the delay. In this paper, we investigate the joint sub-carrier allocation and caching placement for two network slices, and an average delay optimization problem for one slice with user data rate guarantee for the other slice is formulated, which can be solved by a two-step iterative algorithm. The core idea is to optimize decoupled variables iteratively via Hungarian method, linear integer programming, and geometric programming(GP) with the help of decomposition. Simulation results reveal a fast convergence speed and a near-optimal performance of the proposed algorithm.**

***Index Terms*—Network slicing, Fog radio access network, Resource management, Caching placement.**

I. Introduction

Driven by the emergence of a variety of new use cases and business models, the network slicing technique is proposed to improve network flexibility and reduce operating costs. Through splitting the physical network into several virtual networks as network slices, each slice is customized to provide only functions and resources that is necessary for a specific application. Meanwhile, the F-RAN is regarded as a potential network architecture for high throughput, high energy efficiency, high reliability, and low latency [1], and hence network slicing in F-RANs is attractive.

To harvest the potential advantages of network slicing, the effective resource management for network slices has been studied [2]. In [3], the author developed a joint base station assignment, sub-carrier, and power allocation algorithm to maximize the network sum rate in multi-cell virtualized wireless networks under satisfying the minimum rate constraint of each slice. In [4], the massive multi-input-multi-output technology is introduced to enlarge spatial degree of freedom and enhance the sum-utility of all network slices by optimizing the antennas allocation together with power and sub-carrier allocation. To achieve the fairness among network slices, a criterion based on a weighted proportionally fair object for dynamic resource allocation is proposed in [5]. In addition to the centralized resource management, some researches based on the hierarchical architecture are conducted as well. A two-level hierarchical resource allocation problem is modeled as an auction game in [6], where the network operator is responsible for resource management among network slices, and each slice allocates the resources to its own users.

Generally, majority of previous works focus only on the radio resource allocation for network slicing. In order to make the network slices work in a more optimized way and provide higher quality-of-service, the edge caching in the radio access network can be fully utilized to alleviate the fronthaul burden and reduce the delay. Due to the strong relationship between radio resource allocation and caching placement, how to design the caching strategy is regarded as a novel important consideration for resource management, which has been seldom studied in the literatures. In this paper, a dynamic resource allocation optimization problem that jointly considers the sub-carrier allocation, user association and caching placement is formulated in F-RANs to minimize average delay for one slice and meanwhile guarantee the user data rate of the other slice. To handle the non-convexity of the problem, a two-step iterative algorithm based on relaxation and convexification is developed, which can approach the optimal solution with lower complexity compared with the exhaustive search.

The rest of this paper is organized as follows. In Section II, the system model and problem formulation are presented. Section III introduces the proposed two-step iterative algorithm. Section IV demonstrates the simulation results and detailed analysis, followed by the conclusion in Section V.

II. System Model and Problem Formulation

A. Scenario Description

As illustrated in Fig. 1, the downlink transmission in F-RAN is considered, in which two typical types of network slices coexist. One is *eMBB slice* which is defined to provide enhanced mobile broadband services, and the other is *URLLC slice* which is responsible for ultra-reliable and low latency communications. Due to different performance requirements of two slices, the F-RAN configures customized communication modes for them by mode selection. The *eMBB slice* is equipped with M single-antenna remote radio heads (RRHs), and all RRHs serve users together with the cooperative communication. The *URLLC slice* is consisted of N single-antenna fog access points (F-APs), which are

978-1-5090-3009-5/17/$31.00 ©2017 IEEE

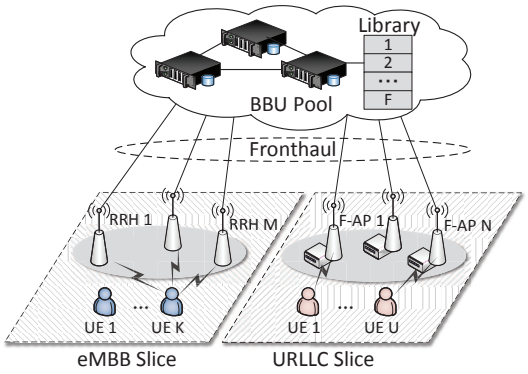

Fig. 1. Two network slices for different application scenarios in fog radio access networks.

equipped with distributed caching, and each F-AP serves part of users independently. Let $\mathcal{M} = \{1, 2, \ldots, M\}$ and $\mathcal{N} = \{1, 2, \ldots, N\}$ denote the set of RRHs and F-APs, respectively. All RRHs and F-APs are connected to a base band unit (BBU) pool with high-capacity fronthaul links. A total of Q single-antenna user equipments (UEs) request to access to the network, among them, K UEs access to *eMBB slice* and the rest U UEs connect with *URLLC slice*. Let $\mathcal{K} = \{1, 2, \ldots, K\}$ and $\mathcal{U} = \{K+1, K+2, \ldots, K+U\}$ denote the set of UEs in two slices, i.e., $Q = K+U$, and the set of all UEs is $\mathcal{Q} = \{1, 2, \ldots, Q\}$. The system bandwidth of W Hz is divided into S sub-carriers and shared by all UEs through orthogonal frequency division multiple access (OFDMA). To ensure each UE can be assigned with at least one sub-carrier, let $\mathcal{S} = \{1, 2, \ldots, S\}$ denote the set of sub-carriers, i.e., $S \geq Q$ and the bandwidth of each sub-carrier is $W_1 = W/S$. Let $h_{i,s,k}$, $g_{j,s,u}$ represent the channel state information on the sub-carrier s from the RRH i to UE k and the F-AP j to UE u, respectively. Each RRH i and F-AP j transmit at a fixed power p_i^R and p_j^F, respectively, and the additive white Gaussian noise is denoted with the distribution $\mathcal{CN}(0, \sigma^2)$.

B. System Model

Due to OFDMA, each sub-carrier cannot be allocated to more than one UE. Let $\alpha_{s,q} \in \{0, 1\}$ be a binary variable, which represents whether the sub-carrier s is allocated to UE q or not. i.e., $\alpha_{s,q} = 1$ implies the sub-carrier s is allocated to UE q, and $\alpha_{s,q} = 0$ otherwise. Thus, the sub-carrier allocation strategy (SCAS) can be defined as follow

$$\boldsymbol{\alpha} = \{\alpha_{s,q} : s \in \mathcal{S}, q \in \mathcal{Q}\} \in \mathbb{C}^{S \times Q} \tag{1}$$

The library of F files, denoted by $\mathcal{F} = \{1, 2, \ldots, F\}$, is stored in the BBU pool. The size of the f-th file is C_f bits and the capacity of j-th F-APs is limited with O_j bits, therefore, each F-AP can only cache part of files. Let binary-valued $\beta_{f,j} \in \{0, 1\}$ represent caching placement indicator, i.e., $\beta_{f,j} = 1$ indicates the file f is cached in F-AP j, and $\beta_{f,j} = 0$ otherwise. The caching placement strategy (CPS) can be given as

$$\boldsymbol{\beta} = \{\beta_{f,j} : f \in \mathcal{F}, j \in \mathcal{N}\} \in \mathbb{C}^{F \times N} \tag{2}$$

Meanwhile, the file preferences of UE u requesting for file f can be defined as the request probability $P_{u,f}$, which are normalized as $\sum_{f=1}^{F} P_{u,f} = 1$.

Considering the different communication modes and performance requirements between network slices, two system models are established, respectively.

1) *eMBB Slice*

In order to provide a higher transmission rate, all RRHs cooperatively transmit the signal to UEs based on the global cloud radio access network mode [7]. Therefore, the total rate of UE k can be expressed as

$$\begin{aligned} R_k &= \sum_{s \in \mathcal{S}} \alpha_{s,k} R_{s,k} \\ &= \sum_{s \in \mathcal{S}} \alpha_{s,k} W_1 \log_2 (1 + \frac{\sum\limits_{i \in \mathcal{M}} |h_{i,s,k}(t)|^2 p_i^R}{\sigma^2}) \end{aligned} \tag{3}$$

where $R_{s,k}$ is the available rate of UE k on sub-carrier s. The performance To ensure basic rate requirements of UEs in *eMBB slice*, each UE k requests for a minimum reserved rate of R_k^{thr}.

2) *URLLC Slice*

In order to shorten the user-perceived delay for delay-sensitive applications, the local F-AP mode [7] is adopted to provide services. Since each UE only can be served by one F-AP and the maximum number of the F-AP j for accessing is limited as A_j, the problem of user association needs to be additional considered to improve network performances. Let $x_{j,u} \in \{0, 1\}$ be a binary variable indicating whether or not UE u is associated with F-AP j, i.e., $x_{j,u} = 1$ indicates the UE u is associated with F-AP j, and $x_{j,u} = 0$ otherwise. The association between UEs and F-APs can be described as the following matrix,

$$\boldsymbol{x} = \{x_{j,u} : j \in \mathcal{N}, u \in \mathcal{U}\} \in \mathbb{C}^{N \times U} \tag{4}$$

When UEs request files from the associated F-APs, the main components of delay are the wireless transmission delay and the fronthaul delay. The transmission delay for transmitting file f from F-AP j to UE u is calculated as

$$D_{j,u,f}^T = \frac{C_f}{R_{j,u}} = \frac{C_f}{\sum\limits_{s \in \mathcal{S}} \alpha_{s,u} R_{j,s,u}} \tag{5}$$

where $R_{j,u}$ is the total transmission rate from F-AP j to UE u, and $R_{j,s,u} = W_1 \log_2 (1 + \frac{|g_{j,s,u}|^2 p_j^F}{\sigma^2})$ is the rate from F-AP j to UE u on sub-carrier s. In practice, the fronthaul delay is related to the average link distance and traffic load. For simplification, we assume the fronthaul delay as a fixed value D^F, which is larger than the transmission delay. The actual

delay will not contain the fronthaul delay if the requested file is cached in the associated F-AP. However, if the file is not cached, the F-AP needs to fetch the file from the BBU pool through fronthaul links first, which produces the fronthaul delay. Thus, the fronthaul delay between UE u and F-AP j for file f can be calculated as

$$D^F_{j,u,f} = (1-\beta_{f,j})D^F \tag{6}$$

Consequently, the delay for UE u which is associated with F-AP j requesting for file f can be written as

$$\begin{aligned} D_{j,u,f} &= D^T_{j,u,f} + D^F_{j,u,f} \\ &= \frac{C_f}{R_{j,u}} + (1-\beta_{f,j})D^F \end{aligned} \tag{7}$$

C. Problem Formulation

To optimize the whole system performance while satisfying the fundamental performance of each slice as the prerequisite, the resource allocation and caching placement should be jointly considered owing to the mutual influence on the ultimate performance. In order to deal with two different performance metrics together, a common method is to combine two metrics into a unified optimization objective, such as spectral efficiency. However, in the current situation, the ratio of rate to delay has not been defined yet, which makes the combination meaningless. Therefore, in this paper, we adopt to individually optimize the performance of *URLLC slice*, which guaranteeing the rate requirement of *eMBB slice*. With the consideration of carrier reuse constraint, accessing capability constraint and caching capacity constraint, the joint resource allocation, user association and caching placement problem to minimize the average delay of UEs in *URLLC slice* is formulated as

$$\min_{\{\boldsymbol{x},\boldsymbol{\alpha},\boldsymbol{\beta}\}} \bar{D} = \frac{1}{U}\sum_{u\in\mathcal{U}}\sum_{f\in\mathcal{F}}\sum_{j\in\mathcal{N}} x_{j,u}P_{u,f}D_{j,u,f} \tag{8}$$

$$\begin{aligned} \text{s.t.} \quad & \text{C1}: \sum_{q\in\mathcal{Q}} \alpha_{s,q} \le 1, \quad \forall s \in \mathcal{S} \\ & \text{C2}: \sum_{f\in\mathcal{F}} \beta_{f,j} C_f \le O_j, \quad \forall j \in \mathcal{N} \\ & \text{C3}: \sum_{j\in\mathcal{N}} x_{j,u} = 1, \quad \forall u \in \mathcal{U} \\ & \text{C4}: \sum_{u\in\mathcal{U}} x_{j,u} \le A_j, \quad \forall j \in \mathcal{N} \\ & \text{C5}: R_k \ge R_k^{thr}, \quad \forall k \in \mathcal{K} \end{aligned}$$

where the optimization objective $\bar{D}$ is the average delay of all UEs requesting any files in *URLLC slice*. C1 is the sub-carrier constraint which ensures each sub-carrier cannot be allocated to more than one UE at the same time. C2 denotes the file caching constraint of each F-AP. C3 and C4 indicate the association constraint of users and F-APs, respectively. C5 is the performance constraint of each UE in *eMBB slice*.

Note that the optimization problem (8) is a non-linear combination optimization problem, which can be proved as a NP-hard problem [8], and it is difficult to find the globally optimal solution. Therefore, proposing an efficient algorithm with reasonable computational complexity is desirable.

III. Two-Step Iterative Algorithm

To tackle the computational complexity of (8), an effective iterative algorithm to optimize the sub-carrier allocation, user association and caching placement for network slices is proposed in this section. The optimization variables are tightly coupled in the objective function, which causes the problem hard to solve directly. Therefore, variable decoupling is applied to reduce the complexity. Firstly, the original problem is decomposed into two sub-problems with independent variables. When each sub-problem is optimized for the optimal variable value, the remaining variables are assumed to be given. Based on the loop iteration, all variables will tend to be stable, and these values can be approximated as optimal solutions to the problem.

A. Joint User Association and Caching Placement Problem

Based on a given SCAS, the transmission delay $D^T_{j,u,f}$ can be regarded as a fixed value. However, the user association variable and caching placement variable are still coupled in the optimization problem. To conquer this challenge, a new variable $z^u_{j,f} = x_{j,u}(1-\beta_{f,j})$ is introduced and the optimization problem can be rewritten as

$$\min_{\{\boldsymbol{x},\boldsymbol{\beta},\boldsymbol{z}\}} \frac{1}{U}\sum_{u\in\mathcal{U}}\sum_{f\in\mathcal{F}}\sum_{j\in\mathcal{N}} P_{u,f}(x_{j,u}D^T_{j,u,f} + z_{j,u,f}D^F) \tag{9}$$

$$\begin{aligned} \text{s.t.} \quad & \text{C2}, \text{C3}, \text{C4} \\ & \text{C6}: z^u_{j,f} = x_{j,u}(1-\beta_{f,j}) \end{aligned}$$

To deal with the problem (9) with non-convex constraint C6, the McCormick convex relaxation is applied to equivalently transform the equality constraint C6 to a series of inequality constraints by using McCormick envelopes [9], which is given by

$$z^u_{j,f} \ge x_{j,u} - \beta_{f,j}, \quad \forall j \in \mathcal{N}, \forall u \in \mathcal{U}, \forall f \in \mathcal{F} \tag{10}$$

$$z^u_{j,f} \le x_{j,u}, \quad \forall j \in \mathcal{N}, \forall u \in \mathcal{U}, \forall f \in \mathcal{F} \tag{11}$$

$$z^u_{j,f} \le 1 - \beta_{f,j}, \quad \forall j \in \mathcal{N}, \forall u \in \mathcal{U}, \forall f \in \mathcal{F} \tag{12}$$

$$0 \le z^u_{j,f} \le 1, \quad \forall j \in \mathcal{N}, \forall u \in \mathcal{U}, \forall f \in \mathcal{F} \tag{13}$$

In order to solve the new optimization problem, we use the Lagrange partial relaxation to relax the constraints (10)-(12), and define the respective set of dual Lagrange multipliers as

$$\varphi^u_{j,f} \ge 0, \quad \forall j \in \mathcal{N}, \forall u \in \mathcal{U}, \forall f \in \mathcal{F} \tag{14}$$

$$\eta^u_{j,f} \ge 0, \quad \forall j \in \mathcal{N}, \forall u \in \mathcal{U}, \forall f \in \mathcal{F} \tag{15}$$

$$\chi^u_{j,f} \ge 0, \quad \forall j \in \mathcal{N}, \forall u \in \mathcal{U}, \forall f \in \mathcal{F} \tag{16}$$

Hence, the Lagrange function is expressed as

$$\begin{aligned} L(\boldsymbol{\varphi},\boldsymbol{\eta},\boldsymbol{\chi},\boldsymbol{x},\boldsymbol{\beta},\boldsymbol{z}) = \frac{1}{U}\sum_{u\in\mathcal{U}}\sum_{f\in\mathcal{F}}\sum_{j\in\mathcal{N}}\Big[& P_{u,f}x_{j,u}D^T_{j,u,f} \\ & +P_{u,f}z_{j,u,f}D^F+\varphi^u_{j,f}(x_{j,u}-\beta_{f,j}-z^u_{j,f}) \\ & +\eta^u_{j,f}(z^u_{j,f}-x_{j,u})+\chi^u_{j,f}(z^u_{j,f}+\beta_{f,j}-1)\Big] \\ & = f(\boldsymbol{x})+f(\boldsymbol{\beta})+f(\boldsymbol{z}) \end{aligned} \tag{17}$$

where $f(\boldsymbol{x})$, $f(\boldsymbol{\beta})$ and $f(\boldsymbol{z})$ are the polynomial functions with corresponding variables. Thus, the dual problem can be reformulated as

$$\begin{aligned} & \max_{\{\boldsymbol{\varphi},\boldsymbol{\eta},\boldsymbol{\chi}\}}\min_{\{\boldsymbol{x},\boldsymbol{\beta},\boldsymbol{z}\}} L(\boldsymbol{\varphi},\boldsymbol{\eta},\boldsymbol{\chi},\boldsymbol{x},\boldsymbol{\beta},\boldsymbol{z}) \\ & \text{s.t.} \quad \text{C2},\text{C3},\text{C4},(13)-(16) \end{aligned} \tag{18}$$

The dual problem (18) can be decomposed into three sub-problems with independent feasible region, respectively, which can be rewritten as

$$\begin{aligned} \text{P1}: & \min_{\{\boldsymbol{x}\}}\sum_{u\in\mathcal{U}}\sum_{f\in\mathcal{F}}\sum_{j\in\mathcal{N}} x_{j,u}(P_{u,f}D^T_{j,u,f}+\varphi^u_{j,f}-\eta^u_{j,f}) \\ & \text{s.t.} \quad \text{C3},\text{C4} \\ \text{P2}: & \min_{\{\boldsymbol{\beta}\}}\sum_{u\in\mathcal{U}}\sum_{f\in\mathcal{F}}\sum_{j\in\mathcal{N}} \beta_{f,j}(\chi^u_{j,f}-\varphi^u_{j,f}) \\ & \text{s.t.} \quad \text{C2} \\ \text{P3}: & \min_{\{\boldsymbol{z}\}}\sum_{u\in\mathcal{U}}\sum_{f\in\mathcal{F}}\sum_{j\in\mathcal{N}} z^u_{j,f}(P_{u,f}D^F-\varphi^u_{j,f}+\eta^u_{j,f}+\chi^u_{j,f}) \\ & \text{s.t.} \quad (13) \end{aligned} \tag{19}$$

Particulary, after the decomposition, the joint optimization problem becomes essentially separate optimization problems. The sub-problem P1 is a typical assignment problem, which can be solved by Hungarian method. And both sub-problem P2 and P3 are the liner integer optimization problems, which can be solved by the generic linear integer programming method.

By solving the sub-problems and obtaining the values of $\boldsymbol{x}$, $\boldsymbol{\beta}$, $\boldsymbol{z}$, the sub-gradient method is used to update the dual variables. In the iteration t_1, the dual variables are updated as follow:

$$\varphi^u_{j,f}(t_1+1) = [\varphi^u_{j,f}(t_1)+\delta(t_1)d(\varphi^u_{j,f}(t_1))]^+ \tag{20}$$

$$\eta^u_{j,f}(t_1+1) = [\eta^u_{j,f}(t_1)+\delta(t_1)d(\eta^u_{j,f}(t_1))]^+ \tag{21}$$

$$\chi^u_{j,f}(t_1+1) = [\chi^u_{j,f}(t_1)+\delta(t_1)d(\chi^u_{j,f}(t_1))]^+ \tag{22}$$

where $[x]^+ = \max\{0,x\}$ and $\delta(t_1)$ is the step size. And $d(\varphi^u_{j,f}(t_1))$, $d(\eta^u_{j,f}(t_1))$, $d(\chi^u_{j,f}(t_1))$ are the sub-gradients of dual problems, which is expressed as

$$d(\varphi^u_{j,f}(t_1)) = x_{j,u}-\beta_{f,j}-z_{j,f} \tag{23}$$

$$d(\eta^u_{j,f}(t_1)) = z^u_{j,f}-x_{j,u} \tag{24}$$

$$d(\chi^u_{j,f}(t_1)) = z^u_{j,f}+\beta_{f,j}-1 \tag{25}$$

Based on the proof in [10], the problem (9) is guaranteed to converge to optimal values, which can be considered as the best user association and CPS. The concrete implementation algorithm of this process is summarized in Algorithm 1. Step A.

B. Sub-carrier Allocation Problem

With the derived user association and CPS, the optimization problem (8) can be simplified to only solve the sub-carrier allocation problem with single variable. Thus, the optimization problem is rewritten as

$$\begin{aligned} & \min_{\{\boldsymbol{\alpha}\}}\sum_{u\in\mathcal{U}}\sum_{f\in\mathcal{F}}\sum_{j\in\mathcal{N}}\frac{H_{j,u,f}}{R_{j,u}} \\ & \text{s.t.} \quad \text{C1},\text{C5} \end{aligned} \tag{26}$$

where $H_{j,u,f} = P_{u,f}x_{j,u}C_f$ is a known and fixed value in each iteration. In order to reduce the complexity of carrier allocation between two slices, the UEs in *eMBB slice* have the priority to be allocated proper sub-carriers to meet the rate constraint C5. Then, the set of unallocated carriers $\mathcal{S}'$ is reallocated to UEs in *URLLC slice*. Thus, the original constraint C1 can be rewritten as

$$\text{C1}': \sum_{u\in\mathcal{U}}\alpha_{s,u}\le 1, \quad \forall s\in\mathcal{S}' \tag{27}$$

Since the optimization problem (26) is non-linear with high complexity, the GP can be used to solve such problem effectively. Firstly, relax the sub-carrier allocation variable to be continuous as $\alpha_{s,u}\in[0,1]$. Then convert optimization function into the standard GP formulation [11] with the help of successive convex approximation (SCA) [12]. The arithmetic-geometric mean approximation (AGMA) can be applied as the method of SCA to transform the posynomial function into the monomial form.

With the help of AGMA, the posynomial denominator of (26) can be approximated as the product of monomial functions. In iteration t_2, the rate $R_{j,u} = \sum_{s\in\mathcal{S}'}\alpha_{s,u}R_{j,s,u}$ can be approximated as

$$\widetilde{R}_{j,u}(t_2) = \prod_{s\in\mathcal{S}'}\left(\frac{\alpha_{s,u}(t_2)R_{j,s,u}}{\tau_{j,s,u}(t_2)}\right)^{\tau_{j,s,u}(t_2)} \tag{28}$$

where $\tau_{j,s,u}(t_2)$ is expressed as

$$\tau_{j,s,u}(t_2) = \frac{\alpha_{s,u}(t_2-1)R_{j,s,u}}{\sum_{s\in\mathcal{S}'}\alpha_{s,u}(t_2-1)R_{j,s,u}} \tag{29}$$

Therefore, the sub-carrier association problem (26) at each iteration t_2 can be transformed into the following standard GP problem,

$$\begin{aligned} & \min_{\{\boldsymbol{\alpha}\}}\sum_{u\in\mathcal{U}}\sum_{f\in\mathcal{F}}\sum_{j\in\mathcal{N}} H_{j,u,f}\widetilde{R}_{j,u}(t_2)^{-1} \\ & \text{s.t.} \quad \text{C1}' \end{aligned} \tag{30}$$

The problem (30) can be solved by available software packages, e.g., CVX. The algorithm for solving this sub-problem is described in Algorithm 1. Step B.

In order to solve the optimization problem, the two-step iterative algorithm is executed to alternately optimize, which will not stop until the differences between solutions in each iteration are smaller than the threshold. That can guarantee the convergence of the proposed Algorithm 1.

Algorithm 1 Two-Step Iterative Algorithm

1: **Global initialization:** Set $t = 0$, $q(t) = 0$ and choose an initial SCAS as the current $\boldsymbol{\alpha}(t)$.
2: **Repeat:** Set $t = t + 1$.
3: **Step A. User Association and Caching Placement:**
4: **Local initialization:** Set $t_1 = 0$, $\varphi^u_{j,f}(t_1) = 0$, $\eta^u_{j,f}(t_1) = 0$, $\chi^u_{j,f}(t_1) = 0$, $q(t_1) = q(t)$, $\boldsymbol{\alpha}(t_1) = \boldsymbol{\alpha}(t)$.
5: **Repeat:**
6: **Step A.1:** Find the solutions of $\boldsymbol{x}, \boldsymbol{\beta}, \boldsymbol{z}$ using (19) and set $t_1 = t_1 + 1$.
7: **Step A.2:** Set $q(t_1) = L(\boldsymbol{\varphi}, \boldsymbol{\eta}, \boldsymbol{\chi}, \boldsymbol{x}, \boldsymbol{\beta}, \boldsymbol{z})$ and update the dual variable $\varphi^u_{j,f}$, $\eta^u_{j,f}$, $\chi^u_{j,f}$ using (20),(21),(22).
8: **Until:** $|q(t_1) - q(t_1 - 1)| \leq \varepsilon_1$, where $0 < \varepsilon_1 \ll 1$. Set $\boldsymbol{x}(t) = \boldsymbol{x}(t_1 - 1)$, $\boldsymbol{\beta}(t) = \boldsymbol{\beta}(t_1 - 1)$ and $q(t) = q(t_1)$.
9: **Step B. Sub-carrier Allocation:**
10: **Local initialization:** Set $t_2 = 0$, $\boldsymbol{x}(t_2) = \boldsymbol{x}(t)$.
11: **Repeat:** Set $t_2 = t_2 + 1$.
12: **Step B.1:** Update $\tau_{j,s,u}(t_2)$ using (29).
13: **Step B.2:** Find optimal SCAS using (30).
14: **Until:** $\left\|\boldsymbol{\alpha}(t_2) - \boldsymbol{\alpha}(t_2 - 1)\right\| \leq \varepsilon_2$, where $0 < \varepsilon_2 \ll 1$. Set $\boldsymbol{\alpha}(t) = \boldsymbol{\alpha}(t_2)$.
15: **Until:** $|q(t) - q(t-1)| \leq \varepsilon_1$ and $\left\|\boldsymbol{\alpha}(t) - \boldsymbol{\alpha}(t-1)\right\| \leq \varepsilon_2$.

C. The Complexity of the Proposed Algorithm

In Step A, to guarantee the accuracy ε_1 of sub-gradient method, this process needs $O(1/{\varepsilon_1}^2)$ iterations, and the time complexity in each iteration is $O(U^3)$, where U denotes the number of users. Therefore, the complexity of the Step A is $O(U^3/{\varepsilon_1}^2)$. In Step B, since CVX is used to solve GP problem with the interior point method, the number of required iteration is $O(\frac{\log(c/(t^0\varepsilon_2))}{\log(\xi)})$, where c is the total number of constraints in (30), t^0 is the initial point to approximate the accuracy of interior point method, ε_2 is the stopping criterion and ξ is used for updating the accuracy of interior point method. The time complexity of Step B required to convert the non-convex problem to (30) using AGMA is $O(UFNS'+US')$. Therefore, the complexity of the Step B is $O((UFNS'+US')\times(\frac{\log(c/(t^0\varepsilon_2))}{\log(\xi)}))$.

IV. Simulation Results and Analysis

A small-scale F-RAN system with two slices is considered during simulations. There are 5 RRHs and 5 UEs in *eMBB slice*, and 3 F-APs and 6 UEs in *URLLC slice*, respectively. The number of sub-carriers is $S = 20$ and the bandwidth of each sub-carrier is $W_1 = 15$KHz. There are a total of 10 files and the size of each file is 1 Kbits. Each F-AP can cache 2 files and allow 3 UEs to access in. The transmission powers are set by $p_i^R = 8$W and $p_j^F = 10$W, respectively. The fronthaul delay is 0.5s. The reserved rate of UEs is assumed as 8 bps/Hz.

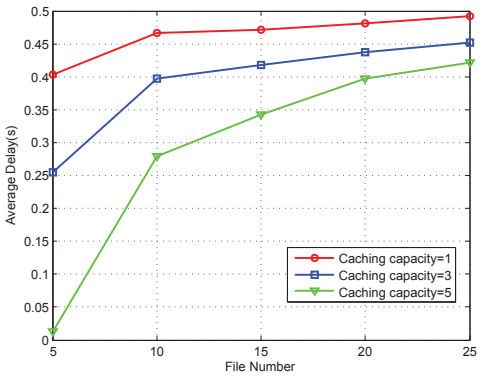

Fig. 2. The effect of caching capacity on the system performance.

Due to the caching capability of F-APs, UEs can get cached files directly from associated F-AP without the fronthaul delay. Fig. 2 approves that the average delay will reduce as the caching capacity of F-APs enhances. However, when the total of files is much larger than the capacity, the trend of decrease is not obvious. Fig. 3 shows that the abundant sub-carries is beneficial to the delay performance, too. With the increase of the available sub-carriers, the average delay decreases gradually. To provide higher reserved rate for UEs in *eMBB slice*, the network will allocate more sub-carriers to these UEs in priority which damages the performances of *URLLC slice*. The distribution of the file request probability and fronthaul delay also have effect on the average delay. Fig. 4 reveals that the performance bring by Zipf distribution is better than that bring by random distribution. As the Zipf parameter increases, the average delay decreases. Meanwhile, with the increase of fronthaul delay, the average delay will increase accordingly.

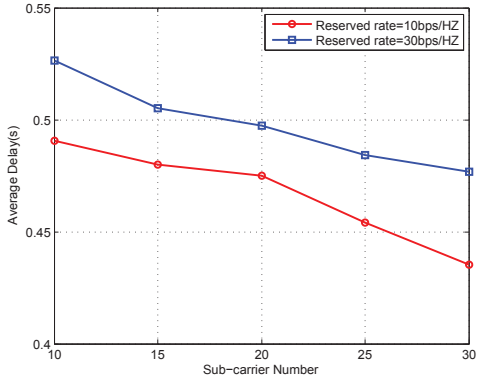

Fig. 3. The effect of sub-carriers on the system performance.

In order to verify the superiority of the proposed algorithm, we compare the performance of the proposal with the exhaustive search which can optimize to optimum solutions

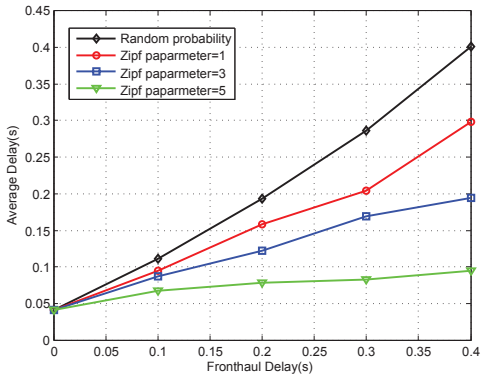

Fig. 4. The effect of file distribution on the system performance.

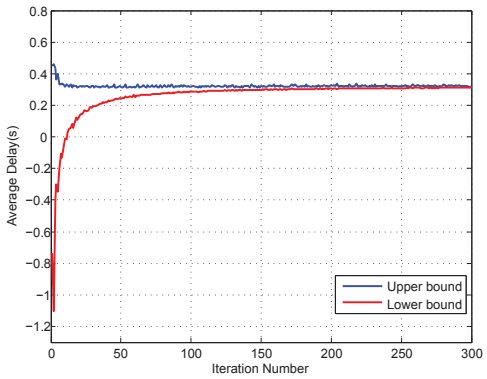

Fig. 6. The convergence of the proposed algorithm.

and achieve the best performance at the cost of high computational complexity. Fig. 5 shows that the performance of the proposal is very close to that obtained using the exhaustive search, and the gap between two algorithms is only about 0.02s.

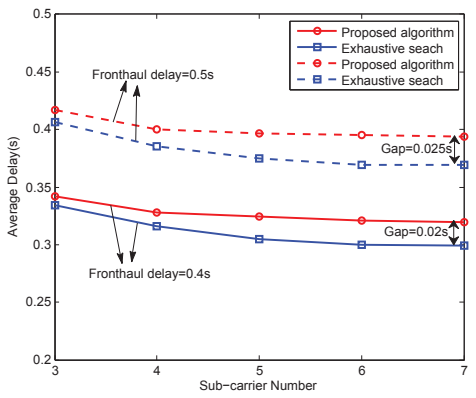

Fig. 5. Performance comparison of the proposed algorithm and exhaustive search.

Meanwhile, the convergence of the Algorithm 1. Step A is presented in Fig. 6. The upper bound is a feasible solution to the primary problem, and the lower bound is a practical solution to the dual problem. It shows that the algorithm converges rapidly in less than a few hundreds iterations with a small error.

V. Conclusion

In this paper, we jointly consider resource allocation and caching placement for two network slices in the F-RAN. An average delay optimization problem considering user data guarantee is formulated, which is non-convex and NP-hard. In order to reduce the complexity, the distributed two-step iterative algorithm is proposed by decomposing the original problem into several sub-problems with different variables, which are solved by Hungarian method, linear programming and GP, respectively. Simulation results reveal that the proposal converges with a fast speed and achieves a near-optimal performance. Therefore, it can be concluded that our work gives a promising method to jointly determine the optimal resource allocation and caching placement strategy.

VI. Acknowledgment

This work was supported in part by the State Major Science and Technology Special Projects (Grant No. 2016ZX03001020-006), the National Natural Science Foundation of China (Grant No.61361166005), the National Basic Research Program of China (973 Program) (Grant No. 2013CB336600), and the National Program for Special Support of Eminent Professionals.

References

[1] M. Peng, and K. Zhang, "Recent advances in fog radio access network: Performance analysis and radio resource allocation," *IEEE Access*, vol. 4, pp. 5003-5009, Aug. 2016.

[2] M. Richart, J. Baliosian, J. Serrat, and J. Gorricho, "Resource slicing in virtual wireless networks: A survey," *IEEE Trans. Netw. Service Manag.*, vol. 13, no. 3, pp. 462-475, Sep. 2016.

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

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

[5] P. Caballero, A. Banchs, G. de Veciana, and X. Costa-P'erez, "Multi-Tenant radio access network slicing: Statistical multiplexing of spatial loads," [Online]. Available: https://arxiv.org/pdf/1607.08271.pdf

[6] 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, Oct. 2016.

[7] 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.

[8] B. Korte, and J. Vygen, *Combinatorial Optimization: Theory and Algorithms*, Berlin, Germany: Springer, 2008.

[9] L. Liberti, and C.C. Pantelides, "An exact reformulation algorithm for large nonconvex NLPs involving bilinear terms," *J.Global Optim.*, vol. 36, no. 2, pp. 161-189, Oct. 2006.

[10] D. P. Bertsekas, *Nonlinear Programming*. Athena Scientific, Belmont, MA, 1995.

[11] G. Xu, "Global optimization of signomial geometric programming problems," *Eur. J. Oper. Res.*, vol. 233, no. 3, pp. 500-510, 2014.

[12] B. R. Marks, and G. P. Wright, "A general inner approximation algorithm for nonconvex mathematical programs," *Oper. Res.*, vol. 24, no. 4, pp. 681-683, 1978.