Energy Efficient User Association, Resource Allocation and Caching Deployment in Fog Radio Access Networks

Xiangnan Liu, Haijun Zhang *Senior Member, IEEE*, Keping Long, *Senior Member, IEEE*, Arumugam Nallanathan, *Fellow, IEEE*, and Victor C. M. Leung, *Fellow, IEEE*

***Abstract*—The heterogeneous fog radio access networks (Fog-RAN), the integration of fog computing, and traditional heterogeneous radio access networks can be implemented through the next-generation wireless communication networks. However, most of the solutions are limited to the spectrum efficiency optimization, and cross-tier interference existing in the fog access points (F-APs) could affect the network performance seriously. In this paper, the user association, resource allocation (including bandwidth and power), and caching deployment are investigated in the heterogeneous Fog-RAN to consider energy efficiency and cross-tier interference mitigation. Specifically, the user association, resource allocation, and caching strategy are formulated as a non-convex optimization problem and then transformed into a convex problem, which can be solved by a proposed algorithm based on the concept of the alternating direction method of multipliers (ADMM). Then an ADMM-based algorithm is proposed to enhance the energy efficiency of the Fog-RAN. Compared with the current solutions, simulation results illustrate the proposed algorithm's convergence and effectiveness.**

***Index Terms*—Fog radio access networks (Fog-RAN), energy efficiency, resource allocation, caching deployment.**

I. Introduction

With the rapid development of the Internet of Things (IoT), many smart devices have exponential growth recently. According to the statistical data from market research firm IDC, the number of global smart devices will reach around 41.6 billion by 2025, generating 79.4ZB data [1]. The loading of backhaul links will surge with billions of the IoT devices and the phenomenons of packet loss in transmission link are easier to occur. To meet future mobile services' complex requirements, the new generation broadband mobile communication with large capacity and high transmission rate requires development and breakthrough. In the traditional heterogeneous networks, the spectral resource and energy resource between low-power access points or base stations (BSs) are easy to be limited [2], which leads to a large gap between the performance gains of coordinated multi-point transmission. Besides, the constrained backhaul quickly causes more complicated network management, high cost, low spectrum efficiency, and low energy efficiency [3].

As an emerging paradigm in the future wireless communication networks, the edge computing enables the cloud computing capabilities to sink to mobile devices. The edge layer between the terminal apparatus and the cloud is implemented differently because the edge layer uses the communication protocol and services. There are three kinds of typical techniques including mobile edge computing, cloudlet computing, and fog computing [4]. The mobile edge computing can bring computing and storage capacity to the boundary of the cloud infrastructure [5]. Compared with the existing frameworks and algorithms for task offloading, mostly focusing on clouds, cloudlets have been quickly gaining recognition as an alternative offloading destination [6]. As for the fog computing, it realizes a set of distributing functions to perform resource allocation, storage management, and computing services [7].

Based on the discussion above, performances among these three techniques are analyzed. As for power consumption, the fog computing consumes lower energy compared to the above two [8]. And when it comes to the applications targeted by these three techniques, the fog computing goes far beyond the cloudlets and the mobile edge computing. The fog computing sinks down a lot of communication and storage capabilities at the edge of networks, extending the current cloud computing paradigm to the smart devices [9].

Thus, we choose the fog radio access networks (Fog-RAN) as the deployment environment to adapt massive data and real-time requirements from millions of sensors [10]. In the Fog-RAN, the powerful capacity of processing and computing sinks in the edge devices, including access points and user equipments. The edge devices undertake some tasks to reduce the backhaul load. The Fog-RAN is an emerging domain that tackle enormous requests from growing user equipments. Beyond the centralized architecture of cloud radio access networks [11], many key functions are distributed to the Fog-RAN's edge. It can provide a superior user experience and improve total network performance. The edge caching are installed in the fog access points (F-APs) [12]. Caching deployment is a problem that looks into how the diversity contents is cached in F-APs because it has significant impact to the network performance of the Fog-RAN. Therefore, how to deploy the caching better in the Fog-RAN has become a

X. Liu, H. Zhang, and K. Long are with Beijing Engineering and Technology Research Center for Convergence Networks and Ubiquitous Services, University of Science and Technology Beijing, Beijing, China, 100083 (email: xiangnan.liu@xs.ustb.edu.cn, haijunzhang@ieee.org, longkeping@ustb.edu.cn).

Arumugam Nallanathan is with the School of Electronic Engineering and Computer Science, Queen Mary, University of London (e-mail: a.nallanathan@qmul.ac.uk)

Victor C. M. Leung is with the Department of Electrical and Computer Engineering, The University of British Columbia, Vancouver, BC V6T 1Z4 Canada (e-mail: vleung@ece.ubc.ca).

focus [13], [14]. As one of the most important parts of the Fog-RAN, a trade-off between the transmission bandwidth cost and the storage cost is the design principle of caching deployment [15], [16]. However, the scale of content acquired by content providers is growing significantly, and it is thus all but impossible to cache all content [17], [18]. The fog user equipments (F-UEs) can access the interesting information within one hop, which significantly reduces the latency. It is necessary to study the content delivery performance in caching-aided heterogeneous wireless networks to illustrate the benefits of placing caching distributed over the whole Fog-RAN.

A. Motivations and Contributions

According to the above introduction, the Fog-RAN is an advanced technique for the future wireless communication system which can provide high spectral efficiency [19]. And the problem of achieving maximum energy efficiency to promote the system performance of the Fog-RAN is of utmost importance. How to realize the trade-off between circuit power consumption to sustain caching deployment and transmit power consumption without caching deployment is worthy of study.

Meanwhile, most of the studies are limited to a single field among these aspects, while relatively less on joint optimization in all these aspects. There are a few reviews on user association and resource allocation in wireless communication networks. In [20], Abedin et al. proposed a two-sided matching game to realize the optimization of user association and bandwidth allocation for typical IoT applications in the Fog-RAN. Qi et al. investigate a stochastic-geometry approach to user association and time-frequency resource block allocation in the NOMA based Fog-RAN [21]. The joint optimization of user association, bandwidth, power allocation, and caching deployment simultaneously in the heterogeneous Fog-RAN is seldom mentioned.

The edge caching gain is usually calculated by two indexes, including time delay reduction [22] and released bandwidth [23]. Compared with the released bandwidth, the consideration of time delay reduction is feeble. Network latency is a key performance index that affects the quality of the user experience. Reducing latency is a crucial indicator to measure caching performance simultaneously, so it is also considered in this paper.

Additionally, the cross-tier interference in the spectrum-sharing deployment of F-APs could affect the network performance seriously [24]. This impact can not be ignored in the heterogeneous Fog-RAN. Effective cross-tier interference management ensures stable coexistence between MBS and F-APs in the heterogeneous Fog-RAN, worthy of being studied.

Motivated by these observations, the preliminary investigation on this research problem was published in [25], and this work extends in the following ways: (1) The power allocation changes into joint optimization with bandwidth allocation and caching deployment; (2) The minimum transmit data rate constraint is supplied to guarantee the Quality of Service (QoS) for F-UEs; (3) We reformulate the caching gain added with the consideration of time delay; (4) More simulation results are provided to verify the effectiveness of the proposed algorithm.

In this paper, the environment mainly depends on the Fog-RAN, considering heterogeneous networks and QoS. User association, wireless resource allocation (bandwidth and power), caching deployment with energy efficiency are considered in the heterogeneous Fog-RAN. We formulate user association, resource allocation (bandwidth and power), and caching deployment as a joint optimization problem. The caching gains both considering alleviation of bandwidth and time delay reduction are taken into consideration in the proposed networks architecture. Simulation results illustrate the superior performance of the proposed algorithm.

The main contributions of this paper can be summarized as follow:

- The joint optimization problem for a heterogeneous Fog-RAN is formulated, combined with user association, resource allocation, and caching deployment. The energy efficiency maximization with multiple constraints is designed.
- Every F-AP is equipped with edge caching. Gains from edge caching includes the alleviation of backhaul bandwidth and the reduction of time delay. In this paper, both of them are considered in the heterogeneous Fog-RAN with energy efficiency maximum.
- Concerning that the original optimization problem is non-convex and nonlinear, an approximate convex transformation is proposed. This method can convert the objective function into a global consensus problem which is convenient to the ADMM's solving.
- A joint optimization scheme based on the ADMM is proposed for energy efficiency maximization. The effect of this distributed optimization can be optimized by iteration, and the simulation results verify the convergence of the proposed algorithm.

B. Organization

The rest of the paper is organized as follows. Section II presents the heterogeneous Fog-RAN system model and user association problem formulation, wireless resource allocation, and edge caching deployment. Section III provides the algorithm based on ADMM, which aims to achieve user association, resource allocation combined with caching deployment in the heterogeneous Fog-RAN. Section IV clarifies the computational complexity of the proposed algorithm. In Section V, the proposed algorithm is verified by simulations. Finally, the paper is summarized in Section VI.

II. System Model and Problem Formulation

A. System Model

The heterogeneous Fog-RAN consists of one MBS and $\mathcal{J}$ F-APs where j is used to indicate the jth F-AP. Both of them are equipped with a few caches, whose sizes depend on their storage and computing capability. Meanwhile, a cloud data center has a more powerful capability, where MBS and F-APs

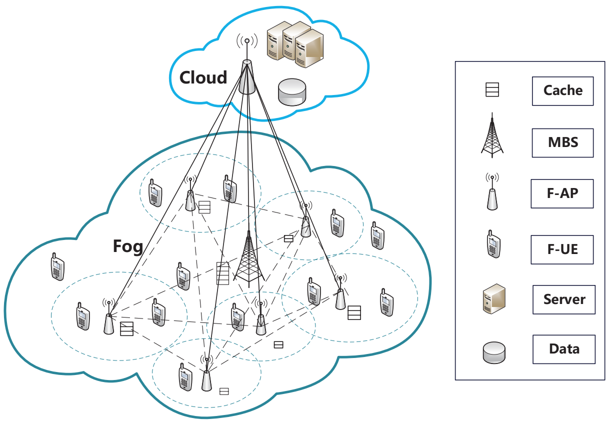

Fig. 1. Heterogeneous Fog-RAN network.

are connected to the cloud data center through backhaul links. $\mathcal{K}$ is denoted as the set of these F-UEs and k is the index of the associated F-UE. There exists co-channel interference between different F-APs. Besides, cross-tier interference from the MBS should also not be ignored.

In Fig. 1, each assigned F-UE is represented by k, and $\mathcal{K}$ is the set of these F-UEs. The spectrum used by different F-APs is shared, which means that there exists co-channel interference between F-APs. Additionally, the spectrum is shared between the MBS and F-APs in the MBS's coverage, so cross-tier interference from the MBS should also be considered. The transmit power on the MBS is $p_{0,k}$ Watts and the F-AP j's is $p_{j,k}$ Watts. Thus, spectrum efficiency of the F-UE k is associated with the F-AP j, which can be calculated as which

$$\gamma_{j,k} = \frac{g_{j,k} p_{j,k}}{\sum\limits_{l,l \neq j} g_{l,k} p_{l,k} + \sigma^2}, \tag{1}$$

where $\gamma_{j,k}$ is the signal-to-interference-plus-noise ratio (SINR) of the F-UE k in the F-AP j. $g_{j,k}$ is the channel gain that includes pathloss and shadowing. σ^2 is the additive white Gaussian noise.

In the design model, let $a_{j,k}$ represent the user association indicator, $a_{j,k} = 1$ when the F-UE k is associated with the F-AP j, otherwise $a_{j,k} = 0$. Actually, each F-UE is connected with only one F-AP, so $\sum\limits_{\mathcal{J}} a_{j,k} = 1$. $b_{j,k} \in [0, 1]$ denotes percentage of spectrum resource allocated and $\sum\limits_{\mathcal{K}} b_{j,k} \leq 1$. The expected information transmission rate of the F-UE k to the F-AP j is:

$$R_{j,k} {=} B_j r_{j,k} b_{j,k}, \tag{2}$$

where B_j is spectrum allocated to the F-AP j. And $r_{j,k} = \log_2 (1 + \gamma_{j,k})$ is the spectrum efficiency of the F-UE k who associates with the F-AP j by using the Shannon bound.

After $R_{j,k}$ from the F-AP j to the F-UE k is obtained, the energy efficiency $\eta_{j,k}$ can be calculated by [26]:

$$\sum_{\mathcal{J}} \sum_{\mathcal{K}} \eta_{j,k} = \frac{B_j r_{j,k} b_{j,k}}{p_{j,k_c} + a_{j,k}(p_{j,k} + m_{j,k} p_{j,k}^{cache})}, \tag{3}$$

where p_{j,k_c} is circuit consumption power, the subscript of $\eta_{j,k}$ is the rule of user association between the F-AP j and the F-UE k. $p_{j,k}^{cache}$ is the power consumption from caching between the F-AP j and the F-UE k. The caching deployment can be controlled by the binary parameter $m_{j,k} \in \{0, 1\}$. If F-AP j caches the requested content of F-UE k, then $m_{j,k} = 1$, otherwise $m_{j,k} = 0$. The first F-UE's index can be used to indicate the content requested by releasing notation.

For convenience, $\mathbf{A}, \mathbf{M}, \mathbf{P}$ is leveraged to replace the part of power consumption:

$$U_p(\mathbf{A}, \mathbf{M}, \mathbf{P}) = p_{j,k_c} + a_{j,k}(p_{j,k} + m_{j,k} p_{j,k}^{cache}), \tag{4}$$

where $U_p(\mathbf{A}, \mathbf{M}, \mathbf{P})$ represents the utility function of the user association matrix $\mathbf{A}$, the caching deployment strategy $\mathbf{M}$, and the power allocation matrix $\mathbf{P}$. When the maximum value of $\eta_{j,k}$ is obtained, the user association matrix $\mathbf{A}$ and the power allocation matrix $\mathbf{P}$ are updated correspondingly. And as for caching deployment strategy $\mathbf{M}$, it will be elaborated in the following.

Moreover, $\zeta(x)$ logarithmic dealing is a non-decreasing and concave function normally. In this paper, the universal logarithmic function is adapted to the utility function. Namely:

$$\zeta(x) = \begin{cases} \log x & x > 0 \\ -\infty & otherwise \end{cases} \tag{5}$$

There are two ways to consider caching gain. One is the alleviation of backhaul bandwidth. The other is the reduction of the time delay. The alleviation of backhaul bandwidth is regarded as the reward of caching in the following formula

$$\Delta C_j^{[1]} = \sum_{\mathcal{K}} q_k \bar{R}_j m_{j,k}, \tag{6}$$

where q_k is the data rate requirement of F-UE k's requested content, $\bar{R}_j$ is the average rate of the single F-UE of F-AP j. $\Delta C_j^{[1]}$ is the estimated reward of caching, which reflects the advantage of considering caching into F-APs.

Concerning the data rate requirement q_k, we regard it follows ZipF distribution. Because caching performance depends on the relative popularity of different objects [22], it has frequently been observed that the popularity of contents follows a generalized ZipF distribution [27], [28], and yields estimates for ε between 0.56 and 0.83. The ε reflects the performance of content reuse, where the most popular files account for the majority of download request.

$$q_k {=} \frac{1/f^{\varepsilon}}{\sum_{f=1}^{F} 1/f^{\varepsilon}}, \forall f. \tag{7}$$

In terms of the reduction of the time delay, the reward of caching $h_{j,k}$ can be obtained:

$$h_{j,k} = \frac{q_k s_k m_{j,k}}{T_{m_{j,k}}}, \tag{8}$$

where $T_{m_{j,k}}$ is the time delay of downloading the requested content through the backhaul link, s_k is the size of the content requested by F-UE k. And through the formula (8), the sum of reward of caching about the time delay can be calculated.

$$\Delta C_j^{[2]} = \sum_{\mathcal{K}} h_{j,k}, \tag{9}$$

where $\Delta C_j^{[2]}$ is the estimated reward of caching which considers the reduction of the time delay.

In this paper, these two situations will be considered together and they can be expressed by

$$\Delta C_j = \Delta C_j^{[1]} + \Delta C_j^{[2]} = \sum_{\mathcal{K}} q_k m_{j,k} \left(\bar{R}_j + \frac{1}{T_{m_{j,k}}} \right). \quad (10)$$

The storage of the F-AP j may be limited and the content size of caching is smaller than the rest space of M_j. It is expressed by:

$$\sum_{\mathcal{K}} a_{j,k} m_{j,k} s_k \leq M_j, \quad (11)$$

and every content is assumed to be equal, i.e. $s_k = 1$.

B. Problem Formulation

In this paper, four important parameters are considered. The user association indicator is $a_{j,k}$, which represents the connection status between the F-UE k and the F-AP j. $b_{j,k}$ and $p_{j,k}$ reflect the allocated bandwidth and power resource status respectively. $m_{j,k}$ indicates updating of caching deployment.

- Firstly, we assume that each F-UE k should be connected with only one F-AP.

$$\sum_{\mathcal{J}} a_{j,k} = 1, \forall k. \quad (12)$$

- The percentage of spectrum resource allocated status should be less than 1.

$$\sum_{\mathcal{K}} b_{j,k} \leq 1, \forall j. \quad (13)$$

- The backhaul bandwidth usage of F-UEs is the same as instantaneous data rate, which is less than the backhaul capacity of F-APs.

$$\sum_{\mathcal{K}} R_{j,k} \leq C_j, \forall j, \quad (14)$$

where $R_{j,k}$ represents expected information transmission rate of the F-UE k to the F-AP j, C_j represents the maximum channel transmission capacity of the F-AP j.

- The caching strategy is limited in the empty space of the caching of each F-AP.

$$\sum_{\mathcal{K}} a_{j,k} m_{j,k} \leq M_j, \forall j, \quad (15)$$

let M_j denote the maximum buffer volume of the F-AP j.

- For each F-UE associated with the F-AP, sum of them exist a power budget $P_{\max}$.

$$\sum_{\mathcal{K}} a_{j,k} p_{j,k} \leq P_{\max}, \forall j, \quad (16)$$

where $P_{\max}$ represents the transmit power threshold.

- The QoS requirement R_k for the F-UE k should maintain its performance [29], which requires the following constraint:

$$\sum_{\mathcal{J}} a_{j,k} B_j r_{j,k} \geq R_k, \forall k. \quad (17)$$

- The cross-tier interference suffered from the MBS to each F-AP j is:

$$\sum_{\mathcal{K}} b_{j,k} p_{j,k} g_{j,k} \leq I_j, \forall j. \quad (18)$$

Let I_j represent the maximum tolerable interference on co-channel for the MBS and each F-AP j.

According to the above discussion of the utility function and constraints, the preliminary problem formulation can be realized as

$$\max_{A,B,M,P} \sum_{\mathcal{J}} \sum_{\mathcal{K}} \frac{B_j r_{j,k} b_{j,k} + q_k m_{j,k} \left(\overline{R}_{j,k} + 1/T_{m_{j,k}} \right)}{U_p(\mathbf{A}, \mathbf{M}, \mathbf{P})}, \quad (19)$$

$$\begin{aligned} \text{s.t.} \quad & C1: \sum_{\mathcal{J}} a_{j,k} = 1, \forall k, \\ & C2: \sum_{\mathcal{K}} b_{j,k} \leq 1, \forall j, \\ & C3: \sum_{\mathcal{K}} R_{j,k} \leq C_j, \forall j, \\ & C4: \sum_{\mathcal{K}} a_{j,k} m_{j,k} \leq M_j, \forall j, \\ & C5: \sum_{\mathcal{K}} a_{j,k} p_{j,k} \leq P_{\max}, \forall j, \\ & C6: \sum_{\mathcal{J}} a_{j,k} B_j r_{j,k} \geq R_k, \forall k \\ & C7: \sum_{\mathcal{K}} b_{j,k} p_{j,k} g_{j,k} \leq I_j, \forall j. \end{aligned} \quad (20)$$

Obviously, optimization problem (19) with constraints (20) is intractable to solve:

- The binary variables $a_{j,k}$ and $m_{j,k}$ are discrete and the set of constraints are non-convex.
- The utility function is not convex due to the product relationship between $a_{j,k}$ and convex function.
- In the heterogeneous Fog-RAN, the number of variables $a_{j,k}$ is very large with the increasing of the F-APs' density.

A mixed discrete and non-convex optimization problem is very difficult to find its global optimal solution. Thus, the optimization problem (19) in the constraints (20) should be simplified.

A relaxation of the binary conditions of $a_{j,k}$ and $m_{j,k}$ constitutes the first step of solving the problem. Following the approach that has been used extensively, $a_{j,k}$ and $m_{j,k}$ are relaxed in the problem (19) to be real value variables, $0 \leq a_{j,k} \leq 1$ and $0 \leq m_{j,k} \leq 1$, where $a_{j,k}$ as well as $m_{j,k}$ can be considered as a time-sharing factor for co-channels and caches. $m_{j,k}$ can be interpreted as the fraction of time that channel is assigned to the F-UE k during one transmission frame with the caching model. The relaxed $a_{j,k}$ can be sensible and meaningfully interpreted as the time-sharing factor representing the ratio of time when the F-UE k associates with the F-AP j.

After relaxing the binary variables, the utility function is still non-convex. Thus, to make the problem (19) tractable and solvable, the second step is necessary. $\widetilde{b}_{j,k}$ and $\widetilde{m}_{j,k}$ are defined as $\widetilde{b}_{j,k} = a_{j,k}b_{j,k}, \widetilde{m}_{j,k} = a_{j,k}m_{j,k}$ respectively.

According to the formula (5), the utility function (19) is transformed into the convex form:

$$\max_{A,B,M,P} \sum_{\mathcal{J}} \sum_{\mathcal{K}} a_{j,k} u \left(\frac{B_j r_{j,k} \widetilde{b}_{j,k} + q_k \widetilde{m}_{j,k} \left(\overline{R}_{j,k} + 1/T_{m_{j,k}} \right)}{a_{j,k} U_p(\mathbf{A}, \mathbf{M}, \mathbf{P})} \right). \tag{21}$$

This is because the $f(t,x) = xlog(t/x), t \geq 0, x \geq 0$ is the well-known perspective operation of log, whose convexity is preserved. The $\log \left(\frac{B_j r_{j,k} \widetilde{b}_{j,k} + q_k \widetilde{m}_{j,k} \left(\overline{R}_{j,k} + 1/T_{m_{j,k}} \right)}{U_p(\mathbf{A},\mathbf{M},\mathbf{P})} \right)$ is concave, so the $a_{j,k} \log \left(\frac{B_j r_{j,k} \widetilde{b}_{j,k} + q_k \widetilde{m}_{j,k} \left(\overline{R}_{j,k} + 1/T_{m_{j,k}} \right)}{a_{j,k} U_p(\mathbf{A},\mathbf{M},\mathbf{P})} \right)$ is also concave with the help of perspective function. Added with the constraints (20) are linear so the problem transforms into a convex problem.

To solve the problem (21), local copies of the global association indicators are also introduced. Each local variable can be considered the preference of each F-AP j with F-UEs' association. A set of new variables to represent the local copies of user association indicators are introduced. To avoid confusion, l denotes the subscript of other F-APs rather than j. The local copy of the total user association at F-AP j can be denoted as $\widehat{a}_{l,k}^{j}$. Formally,

$$\widehat{a}_{l,k}^{j} = a_{l,k}, \forall j. \tag{22}$$

Through this local transformation, the problem (21) turns to a global consensus problem, which contributes to introducing the following algorithm to solve.

For the convenience of readers, the parameters concerning the formulation can be found in Table I:

TABLE I
EXPLANATION OF ABBREVIATIONS.

Notations	Explanations
$\mathcal{J}$	The number of F-APs
$\mathcal{K}$	The number of F-UEs
$g_{j,k}$	The channel gain from F-AP j to F-UE k
$p_{j,k}$	The transmit power from F-AP j to F-UE k
$a_{j,k}$	The user association indicator
$b_{j,k}$	The percentage of wireless resource allocated
$m_{j,k}$	The caching deployment
$r_{j,k}$	The spectrum efficiency
$\eta_{j,k}$	The energy efficiency
$\widetilde{b}_{j,k}$	The relaxed bandwidth allocation parameters
$\widetilde{m}_{j,k}$	The relaxed caching deployment parameters
$U_p(\mathbf{A}, \mathbf{P})$	The power consumption function
q_k	The data rate requirement of F-UE k
R_j	The average rate of single F-UE of F-AP j
$T_{m_{j,k}}$	The time delay of downloading the requested content
$\widehat{a}_{l,k}^{j}$	The local copy of a at F-AP j
$\Delta C_j^{[1]}$	The reward of caching about bandwidth alleviation
$\Delta C_j^{[2]}$	The reward of caching about time delay

III. User Association, Resource Allocation And Caching Deployment By Alternating Direction Method of Multipliers

In this section, the ADMM algorithm is introduced firstly. Then the ADMM with consensus constraint is applied into the optimization problem (21). Finally, the optimization problem (21) will be solved by the ADMM with consensus constraint.

A. Alternating Direction Method of Multipliers with Consensus Constraint

The ADMM is suitable for solving convex optimization problems via breaking them into smaller pieces [30]. Typically, the ADMM is used to solve optimization problems with only equation constraints for two optimization variables. It is widely used in signal processing, image processing, machine learning, engineering computing, and other fields, with fast convergence speed, convergence performance advantages.

The ADMM can be regarded as the cooperation of dual decomposition and augmented Lagrangian methods. It adopts the decomposition method to solve a complex global problem with the solution to small local sub-problems. The global consensus problem can be rewritten with local variables $x_i \in R_n$ and a common global variable z:

$$\begin{array}{l} \min \sum_{i=1}^{N} f_i\left(x_i\right), \\ s.t.\ x_i - z = 0, \quad i = 1, 2, ..., N. \end{array} \tag{23}$$

The ADMM for the problem (23) can be transformed into the augmented Lagrangian:

$$\begin{array}{l} L_\rho\left(x_1, ..., x_N, z, \lambda\right) = \\ \sum_{i=1}^{N} \left(f_i\left(x_i\right) + \lambda_i^T \left(x_i - z\right) + (\rho/2) \left\| x_i - z \right\|_2^2 \right). \end{array} \tag{24}$$

Then an optimal result can be generated by the ADMM algorithm [31]:

$$\begin{array}{l} x_i^{t+1} := \\ \arg\min\limits_{x_i} \left(f_i\left(x_i\right) + \lambda_i^{tT} \left(x_i - z^t\right) + (\rho/2) \left\| x_i - z^t \right\|_2^2 \right), \end{array} \tag{25}$$

$$z_i^{t+1} := \arg\min \sum_{i=1}^{N} \left(\lambda_i^{tT} \left(x_i - z^t\right) + (\rho/2) \left\| x_i - z^t \right\|_2^2 \right), \tag{26}$$

$$\lambda_i^{t+1} := \lambda_i^t + \rho \left(x_i^{t+1} - z^{t+1} \right). \tag{27}$$

B. Solution to the problem by ADMM

In this subsection, the ADMM with consensus constraint is used to solve user association, resource allocation (including bandwidth and power), and caching deployment.

In the simplified optimization problem (21), the utility function should be considered. In order to realize the transformation of the global consensus problem, the solution has to convert the maximum into the minimum. Thus, according to the problem (21) and the formula (22), it should be changed as follows:

$$\max_{A,B,M,P} \sum_{\mathcal{J}} \sum_{\mathcal{K}} \eta_{j,k} = \widehat{a}_{l,k}^{j} \zeta \left(\frac{B_j r_{j,k} \widetilde{b}_{j,k} + q_k \widetilde{m}_{j,k} \left(R_{j,k} + 1/T_{m_{j,k}} \right)}{\widehat{a}_{l,k}^{j} U_p(\mathbf{A},\mathbf{M},\mathbf{P})} \right). \tag{28}$$

So $\zeta_j(\cdot)$ in the formula (5) is introduced to transform the utility function $\eta_{j,k}$ into a concave form, is expressed as:

$$\zeta_j = \begin{cases} \sum_{\mathcal{K}} \eta_{j,k} & \eta_{j,k} \in (20) \\ \infty & \text{otherwise} \end{cases}. \tag{29}$$

Finally, the ζ_j is combined with the constraints (20) in the problem (19), and obtain the convex optimization in global consensus form:

$$\begin{aligned} &\min_{A,B,M,P} U_{total} = -\sum_{j=1}^{\mathcal{J}} \zeta_j, \\ &s.t. \widehat{a}_{l,k}^{j} - a_{l,k} = 0, j = 1, 2, ..., \mathcal{J}, \end{aligned} \tag{30}$$

where U_{total} denotes as the total energy efficiency in the proposed heterogeneous Fog-RAN. According to the joint resource optimization model above, the augmented Lagrange function is:

$$\begin{aligned} L_\rho \left(\left\{ \widehat{a}_{l,k}^{j}, \widetilde{b}_{j,k}, \widetilde{m}_{j,k}, p_{j,k} \right\}, \{a_{l,k}\}, \{\boldsymbol{\lambda}^j\} \right) = U_{total} + \\ \sum_{\mathcal{J}} \sum_{\mathcal{K}} \boldsymbol{\lambda}^j \left(\widehat{a}_{l,k}^{j} - a_{l,k} \right) + \frac{\rho}{2} \sum_{\mathcal{J}} \sum_{\mathcal{K}} \left(\widehat{a}_{l,k}^{j} - a_{l,k} \right)^2, \end{aligned} \tag{31}$$

where the Lagrange parameter is ρ, and Lagrange multipliers are $\boldsymbol{\lambda}$.

The ADMM-based algorithm can be realized in the following three updating steps, including joint optimization with local optimization $\left\{ \widehat{a}_{l,k}^{j}, \widetilde{b}_{j,k}, \widetilde{m}_{j,k}, p_{j,k} \right\}_{j \in \mathcal{J}}^{j(t+1)}$, global user association $\mathbf{a}^{(t+1)}$, and Lagrange multipliers $\boldsymbol{\lambda}^{(t+1)}$.

The updating of joint optimization with local optimization $\left\{ \widehat{a}_{l,k}^{j}, \widetilde{b}_{j,k}, \widetilde{m}_{j,k}, p_{j,k} \right\}_{j \in \mathcal{J}}^{j(t+1)}$ is:

$$\begin{aligned} \left\{ \widehat{a}_{l,k}^{j}, \widetilde{b}_{j,k}, \widetilde{m}_{j,k}, p_{j,k} \right\}_{j \in \mathcal{J}}^{j(t+1)} = \arg\min \{ U_{total} + \\ \sum_{\mathcal{J}} \sum_{\mathcal{K}} \lambda_{l,k}^{j(t)} \left(\widehat{a}_{l,k}^{j(t)} - a_{l,k}^{(t)} \right) + \frac{\rho}{2} \sum_{\mathcal{J}} \sum_{\mathcal{K}} \left(\widehat{a}_{l,k}^{j(t)} - a_{l,k}^{(t)} \right)^2 \right\}, \end{aligned} \tag{32}$$

where the local optimization $\left\{ \widehat{a}_{l,k}^{j}, \widetilde{b}_{j,k}, \widetilde{m}_{j,k}, p_{j,k} \right\}_{j \in \mathcal{J}}^{j(t+1)}$ consists of four parameters: local user association among all F-UEs and other F-APs $\widehat{a}_{l,k}^{j(t)}$, bandwidth resource parameter $\widetilde{b}_{j,k}$, caching deployment $\widetilde{m}_{j,k}$, and power allocation parameter $p_{j,k}$ between the F-AP j and the F-UE k. Additionally, $a_{j \in \mathcal{J}}^{(t)}$ denotes the association between with all F-UEs and the F-AP j when the iteration is the $t+1$th. $\lambda_{l,k}^{j(t)}$ is Lagrange multipliers that associates with the tth constraints, $\widehat{a}_{l,k}^{j(t)}$ denotes as the user association copied information of other F-APs in the F-AP j.

The associated F-UEs' optimization matrix is (33), where $\mathcal{J}$ is denoted the gather of F-APs, the updating of global user association $a^{(t+1)}$ is:

$$\begin{aligned} x_{j \in \mathcal{J}}^{(t+1)} = \arg\min \sum_{\mathcal{J}} \left\{ \sum_{\mathcal{J}} \sum_{\mathcal{K}} \lambda_{l,k}^{j(t)} \left(\widehat{a}_{l,k}^{j(t)} - a_{l,k}^{(t)} \right) \right. \\ \left. + \frac{\rho}{2} \sum_{\mathcal{J}} \sum_{\mathcal{K}} \left(\widehat{a}_{l,k}^{j(t)} - a_{l,k}^{(t)} \right)^2 \right\}. \end{aligned} \tag{33}$$

The updating of Lagrange multipliers $\boldsymbol{\lambda}$ is:

$$\lambda_{k\, j \in \mathcal{J}}^{\;j(t+1)} = \lambda_k^{j(t)} + \rho \left(\widehat{a}_{l,k}^{j(t+1)} - a_{l,k}^{(t+1)} \right). \tag{34}$$

The user associated matrix can be updated by the formula (32), It is difficult to deal with $\left\{ \widehat{a}_{l,k}^{j}, \widetilde{b}_{j,k}, \widetilde{m}_{j,k}, p_{j,k} \right\}_{j \in \mathcal{J}}^{j(t)}$ in the formula (32), so Algorithm 1 based on the logarithmic barrier function interior-point method is considered. The logarithmic barrier function interior-point method forms the next hierarchy [32]. In Algorithm 1, ς is the logarithmic barrier. Given that strict feasible point $\left\{ \widehat{a}_{l,k}^{j}, \widetilde{b}_{j,k}, \widetilde{m}_{j,k}, p_{j,k} \right\}^{(0)}$ in the problem (29), ${u_j}^{(t)}$ can be calculated by $tu_j + \varsigma$. Subsequently, t is updated by the product of μ and the original t. And ς can be expressed by

$$\varsigma = -log(h(\mathbf{x})^T) = -\sum_{i=1}^{m} log(h_i(x)), \tag{35}$$

where m is the number of non-equality constrains, which is 6 in the constraints (20). $h_i(x)$ is the logarithmic form of ith constraint. Then $\left\{ \widehat{a}_{l,k}^{j}, \widetilde{b}_{j,k}, \widetilde{m}_{j,k}, p_{j,k} \right\}_{j \in \mathcal{J}}^{j(t)}$ is used to denote the tth solution to the problem (29). Then the local joint optimal user association, resource allocation, and caching deployment $\zeta_j(\widehat{a}_{l,k}^{j}, \widetilde{b}_{j,k}, \widetilde{m}_{j,k}, p_{j,k})$, and corresponding optimal point $\left\{ \widehat{a}_{l,k}^{j}, \widetilde{b}_{j,k}, \widetilde{m}_{j,k}, p_{j,k} \right\}$ can be obtained until $m/t > \varepsilon$.

Based on the analysis above, Algorithm 1 is summarized below.

Algorithm 1 Logarithmic barrier function interior-point method for $\zeta_j(\widehat{a}_{l,k}^{j}, \widetilde{b}_{j,k}, \widetilde{m}_{j,k}, p_{j,k})$
.

```
Initialize: Given strict feasible point $\left\{ \widehat{a}_{l,k}^{j}, \widetilde{b}_{j,k}, \widetilde{m}_{j,k}, p_{j,k} \right\}^{(0)}$ in the problem (29), the step factor $\mu > 1$, and the threshold error $\varepsilon > 0$.
while $m/t > \varepsilon$.
    a) Compute $t\zeta_j + \varsigma$ to get the ${\zeta_j}^{(t)}$;
    b) Increase $t$: $t = \mu t$.
end while
Output: The local joint optimal user association, resource allocation, and caching deployment $\zeta_j(\widehat{a}_{l,k}^{j}, \widetilde{b}_{j,k}, \widetilde{m}_{j,k}, p_{j,k})$, and corresponding optimal point $\left\{ \widehat{a}_{l,k}^{j}, \widetilde{b}_{j,k}, \widetilde{m}_{j,k}, p_{j,k} \right\}$.
```

We can obtain $\left\{ \widehat{a}_{l,k}^{j}, \widetilde{b}_{j,k}, \widetilde{m}_{j,k}, p_{j,k} \right\}$ and the optimal $\zeta_j(\widehat{a}_{l,k}^{j}, \widetilde{b}_{j,k}, \widetilde{m}_{j,k}, p_{j,k})$ via calculating Algorithm 1. After dealing with $\left\{ \widehat{a}_{l,k}^{j}, \widetilde{b}_{j,k}, \widetilde{m}_{j,k}, p_{j,k} \right\}$, the total energy efficiency U_{total} can be obtained in the proposed system. To solve the transformed optimization problem in the subtractive form (30), Algorithm 2 is proposed.

The problem (30) can be solved via the ADMM's applicability on the global consensus problem. The updated local variables $\left\{\widehat{a}^{j}_{l,k},\widetilde{b}_{j,k},\widetilde{m}_{j,k},p_{j,k}\right\}^{j(t+1)}_{j\in\mathcal{J}}$ and total energy efficiency U_{total} calculated from Algorithm 1. Meanwhile, the updating of local variables in formula (32) is finished. are used in the following steps in updating global $\mathbf{a}^{(t+1)}$ and $\boldsymbol{\lambda}^{(t+1)}$. Finally, the global joint optimization, including user association, resource allocation (bandwidth and power), and caching deployment will be realized in the proposed heterogeneous Fog-RAN. Based on the analysis above, the distributed heterogeneous Fog-RAN algorithm by ADMM can be summarized as Algorithm 2.

Algorithm 2 ADMM-based joint resource optimization in the heterogeneous Fog-RAN.

```
Initialize: Lagrange multiplier $\boldsymbol{\lambda} > 0$ and the stop criterion threshold $\zeta > 0$.
Initialize feasible resource allocation, caching deployment with an equal power distribution $\left\{\widehat{a}^{j}_{l,k},\widetilde{b}_{j,k},\widetilde{m}_{j,k},p_{j,k}\right\}^{(0)}$ and begin with the loop.
for $t = 1$ to $T$ do
  repeat
    a) Determine $\mathbf{a}^{(t)}$ and $\boldsymbol{\lambda}^{(t)}$;
    b) For each F-AP $j$, solve the problem (29) to get optimized $U_{total}$ via Algorithm 1;
    c) Update $\left\{\widehat{a}^{j}_{l,k},\widetilde{b}_{j,k},\widetilde{m}_{j,k},p_{j,k}\right\}^{j(t+1)}_{j\in\mathcal{J}}$ via (32);
    d) Update $\mathbf{a}^{(t+1)}$ by the results of $\left\{\widehat{a}^{j}_{l,k},\widetilde{b}_{j,k},\widetilde{m}_{j,k},p_{j,k}\right\}^{j(t+1)}_{j\in\mathcal{J}}$ via (33);
    e) Update $\boldsymbol{\lambda}^{(t+1)}$ via (34).
  until if $s^{(t+1)}_{dual} = \mathbf{a}^{(t+1)} - \mathbf{a}^{(t)}$ and $\left\|s^{(t+1)}_{j}\right\|_{2} \leq \zeta, \forall j$.
end for
Output:The global joint optimal resource allocation $\left\{a_{j,k},\widetilde{b}_{j,k},\widetilde{m}_{j,k},p_{j,k}\right\}^{j(t+1)}_{j\in\mathcal{J}}$.
```

C. Complexity analysis

In this subsection, the space and time computational complexity of the proposed algorithms are discussed.

In Algorithm 1, the calculation of (29) for F-UEs in each F-AP entails $4\mathcal{JK}$ operations, because there exist four parameters including user association, bandwidth, power allocation and caching deployment $\left\{\widehat{a}^{j}_{l,k},\widetilde{b}_{j,k},\widetilde{m}_{j,k},p_{j,k}\right\}$. In Algorithm 2, the update of local user association $\left\{\widehat{a}^{j}_{l,k},\widetilde{b}_{j,k},\widetilde{m}_{j,k},p_{j,k}\right\}$ in equation (33) entails $4\mathcal{JK}\times\mathcal{J} = 4\mathcal{J}^2\mathcal{K}$ operations because of the local copy of user association. The updating $\mathbf{a}^{(t+1)}$ of equation (33) needs $\mathcal{JK}$ storage, and the updating $\boldsymbol{\lambda}^{(t+1)}$ of equation (34) needs $4\mathcal{JK}$ storage, respectively. Therefore, Algorithm 1's space computational complexity is $O(4\mathcal{JK})$, and Algorithm 2's is $O(4\mathcal{J}^2\mathcal{K}+\mathcal{JK}+4\mathcal{JK}) = O(4\mathcal{J}^2\mathcal{K}+5\mathcal{JK})$. Compared with Algorithm 2, the interior-point method's space computational complexity is $O(\mathcal{J}^{\mathcal{K}})$ and has a much higher complexity.

TABLE II
SUMMARY OF COMPUTATIONAL COMPLEXITY.

Algorithms	Terms
Algorithm 1	Each F-AP performs joint resource allocation locally in (29), here the number of F-APs is $\mathcal{J}$, the number of F-UEs is $\mathcal{K}$. Each F-AP finds $\left\{\widehat{a}^{j}_{l,k},\widetilde{b}_{j,k},\widetilde{m}_{j,k},\widetilde{p}_{j,k}\right\}^{(T)}_{j\in\mathcal{J}}$ satisfied $m/T \leq \varepsilon$.
Algorithm 2	Calculating the complexity of the global consensus problem (30). Updating of $\left\{\widehat{a}^{j}_{l,k},\widetilde{b}_{j,k},\widetilde{m}_{j,k}\widetilde{p}_{j,k}\right\}^{(T)}_{j\in\mathcal{J}}$ requires $O(\mathcal{JT})$ operations and $O(4\mathcal{J}^2\mathcal{K})$ storage. Update of $\mathbf{a}^{(t+1)}$ needs $O(\mathcal{T})$ operations and $O(\mathcal{JK})$ storage. Update of $\boldsymbol{\lambda}^{(t+1)}$ needs $O(\mathcal{JT})$ operations and $O(4\mathcal{JK})$ storage. Deploy I Monte-Carlo simulations to ensure the stability of data.

We assume the joint optimal resource allocation vector can be found within at most T iterations. The time complexity of the proposed Algorithm 1 is modified by $O((T))$ thorough the verification. In Algorithm 2, we need to update each iteration resource allocation, so Algorithm 1 is the substep of proposed Algorithm 2. Thus the total complexity of Algorithm 2 is $O((\mathcal{J}T))$. And to ensure the stability of data, we need to deploy I Monte-Carlo simulations, and the time computational complexity of proposed Algorithm 2 is $O((\mathcal{J}TI))$. Compared with the exponential computational complexity of the exhaustive search, the proposed algorithm has lower complexity. And we have summarized the computational complexity of proposed Algorithm 1 and Algorithm 2 in Table II.

IV. SIMULATION RESULTS AND DISCUSSIONS

The simulation results are discussed and the effectiveness of the proposed algorithm is verified in this section. The position MBS is fixed. And 20 F-APs are randomly distributed in the covered area of the MBS. The radius of the MBS is 500m and the radius of F-APs is 10m. The locations of F-UEs will random changed in the covered are of the fixed F-APs at each shot. After 20 shots, the location of 20 F-APs is changed randomly and run 200 F-UEs shots again. The simulation environment is deployed on an X64 desktop computer equipped with Matlab R2019b.

Fig. 2 illustrates the energy efficiency versus iteration numbers with $\rho = 0.1$, $\rho = 1$, and $\rho = 10$. At the same time, we also introduce the interior-point method and equal power allocation to compare the proposed algorithms' effectiveness. The y-axis is the energy efficiency of different methods, and the x-axis is the iteration step-index. The different effect of parameter ρ in ADMM has different convergence of Algorithm 2. We can clearly identify $\rho = 0.1$ gives a higher rate than $\rho = 1$ and $\rho = 10$, particularly before 20 iterations. It will gradually tend to a fixed value when the number of iterations reaches 30. "Interior-Point method" is a certain algorithm that solves linear and nonlinear convex optimization problems [32]. It can approach the boundary of the feasible set only in the limit. Additionally, as a centralized method, the interior-point method needs higher computing complexity, and the proposed

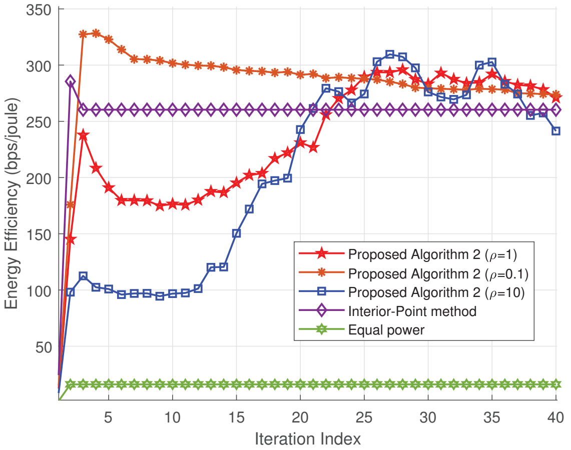

Fig. 2. The energy efficiency convergence under different algorithms.

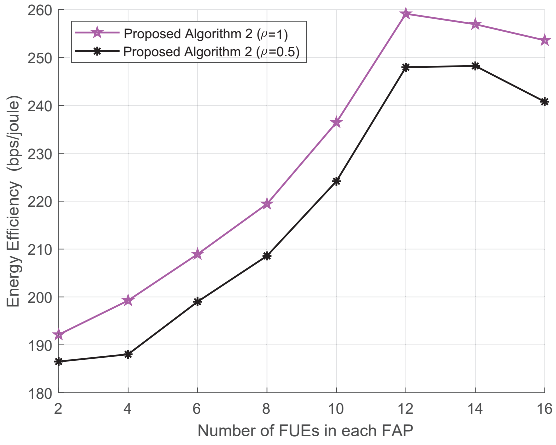

Fig. 4. The effect of proposed Algorithm 2 with the number of F-UEs (energy efficiency).

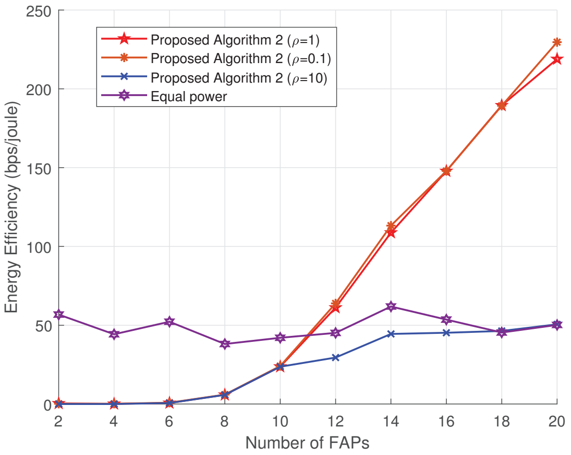

Fig. 3. The effect of proposed Algorithm 2 with the number of F-APs (energy efficiency).

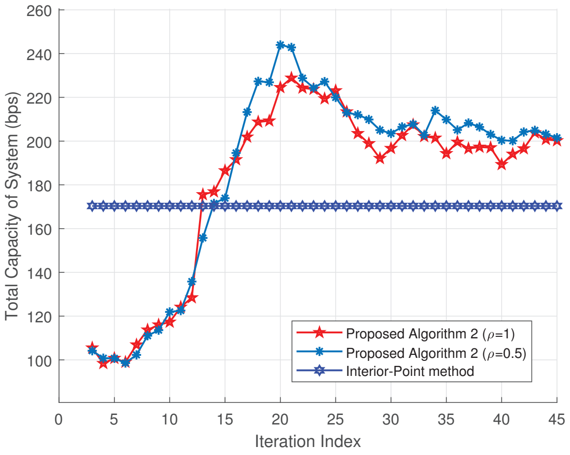

Fig. 5. The total capacity convergence of the proposed Algorithm 2.

algorithm can converge the fixed value that is superior to the interior-point method. And the proposed algorithm can converge the fixed value that is superior to the interior-point method. From Fig. 2, the effect with $\rho = 0.1$ and $\rho = 1$ should be better than $\rho = 10$. On the other hand, the results with $\rho = 0.1$ and $\rho = 1$ will converge the same value with the increasing iteration steps. By simulation result, when $\rho = 0.1$, energy efficiency will have more stable performance, but it will cause unnecessary cost because of low step length. However, $\rho = 1$ also reaches the same convergence when the iteration index is enough. So in the following study, we will study the interval between $\rho = 0.1$ and $\rho = 1$.

Fig. 3 shows energy efficiency with the increasing number of F-APs. $\rho = 1$, $\rho = 0.1$, and $\rho = 0.1$ are chosen to compare. Energy efficiency increasing in $\rho = 0.1$ and $\rho = 1$ is higher than $\rho = 10$, obviously. On the one hand, $\rho = 0.1$ has a slight advantage on the $\rho = 1$, so the size of ρ doesn't matter that so much when it's less than 1.

Fig. 4 describes energy efficiency versus the different F-UEs through ADMM parameters. Obviously, energy efficiency increasing in $\rho = 1$ is higher than $\rho = 0.5$. Fig. 4 shows the energy efficiency of the system when the F-UEs number is from 2 to 16, and the number of F-APs is 20. After reaching 12, the number of F-UEs begins to lose the effectiveness to improve energy efficiency.

Fig. 5 illustrates the convergence of the proposed Algorithm 2 in terms of the system's capacity. In Fig. 5, the iteration index which reaches 40, its converge performance is not very good. The reason is that the proposed objective function is to maximize energy efficiency rather than total capacity. When maximizing the total capacity, it has a longer time to convergence. Therefore, compared with Fig. 2, it needs to iterate to 45 to converge. At this time, the performance is more stable and better.

Fig. 6 depicts the total energy efficiency versus caching constraints. Different cache consumption powers are selected to compare. In the energy efficiency calculation, the tendency is to descend with the increasing of caching consumption power. As is shown in Fig. 7, the energy efficiency of $(power_{cache} = 0.1)$ is significantly superior to others.

To analyze further, the average channel capacity and the

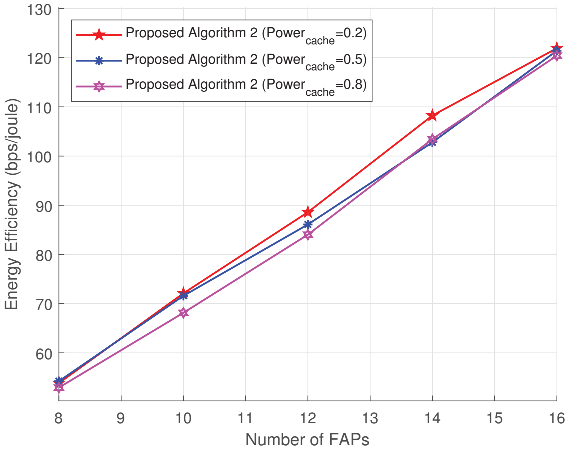

Fig. 6. The energy efficiency with the number of F-APs (caching constraints).

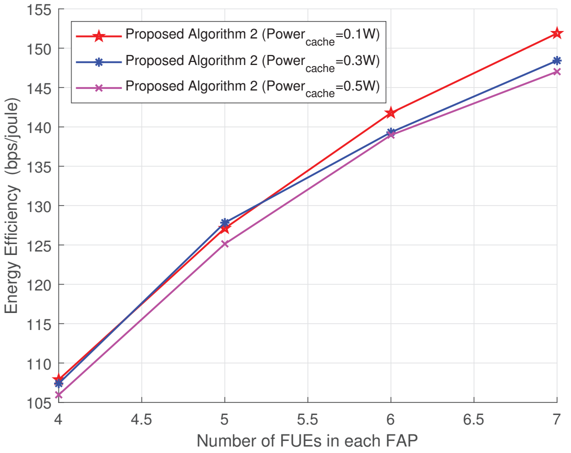

Fig. 7. The average of capacity with the number of F-UEs (caching constraints).

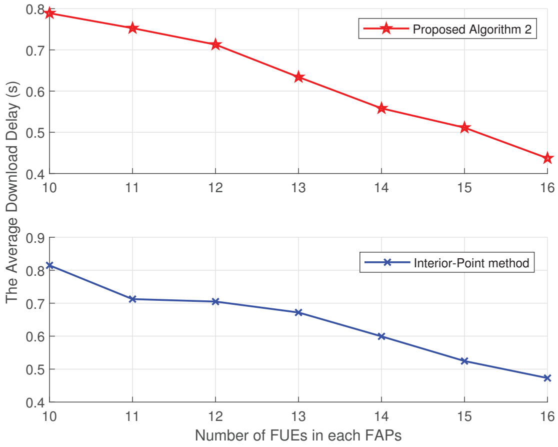

Fig. 8. Comparison of proposed Algorithm 2 and interior-point method (average download delay).

number of F-UEs in each F-AP are shown in Fig. 7, the tendency of energy efficiency is similar to Fig. 6, the case of $(power_{cache} = 0.1)$ is still better than those have a higher power.

Fig. 8 illustrates the average download delay versus different numbers of F-UEs with the proposed Algorithm 2 and the interior-point method. It will gradually tend to a fixed value with the increasing in the number of F-UEs. From Fig. 8, the proposed Algorithm 2 should be better than the interior-point method. This is because the distributed feature of the proposed Algorithm 2 accelerates its local calculating time. Additionally, the proposed Algorithm 2 has a lower space computational complexity than the interior-point method. This suggests the proposed Algorithm 2 is more suitable in this simulation environment in terms of the average download.

Based on the analysis of Fig. 8, Fig. 9 depicts the alleviation of capacity by caching versus the number of F-UEs. The x-axis is the number of F-UEs in each F-AP, and the y-axis is the alleviation of capacity by caching. The alleviating capacity caused by caching is the product of alleviating backhaul bandwidth and transmission rate per unit bandwidth. The proposed Algorithm 2 and the interior-point method continue to be selected to compare. It can be verified that proposed Algorithm 2 has an advantage on the interior-point method with the increasing number of F-UEs. And when the number of F-UEs reaches 10, the alleviation of capacity by caching tends to converge.

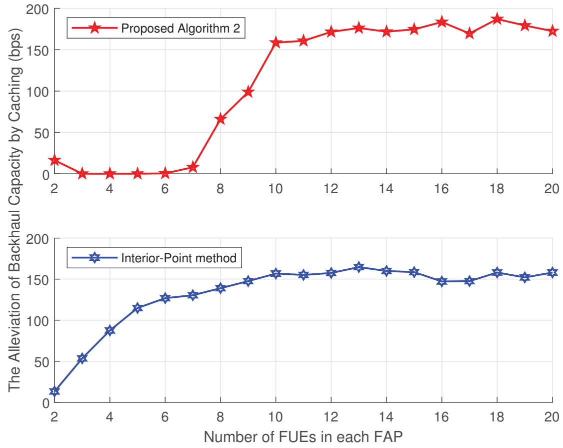

Fig. 9. Comparison of proposed Algorithm 2 and interior-point method (alleviation of capacity by caching).

V. Conclusion

In this paper, we investigated the user association, resource allocation (including bandwidth and power), and caching deployment in the heterogeneous Fog-RAN with the consideration of energy efficiency, capacity, and download delays. The user association, wireless resource allocation, and caching problem were formulated as a global consensus convex optimization problem. An effective algorithm was proposed to solve the energy efficiency maximization in the heterogeneous Fog-RAN using the ADMM method. The effectiveness of the proposed algorithm was verified by simulation results, by comparing it with the current method and changing the proposed algorithm's parameters.

References

[1] M. Chernyshev, Z. Baig, O. Bello, and S. Zeadally, "Internet of Things (IoT): Research, simulators, and testbeds," *IEEE Internet Things J.*, vol. 5, no. 3, pp. 1637–1647, Jun. 2018.

[2] H. Zhang, N. Liu, K. Long, J. Cheng, V. C. M. Leung, and L. Hanzo, "Energy efficient subchannel and power allocation for the software defined heterogeneous VLC and RF networks," *IEEE J. Sel. Areas Commun.*, vol. 36, no. 3, pp. 658–670, Mar. 2018.

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

[4] K. Dolui and S. K. Datta, "Comparison of edge computing implementations: Fog computing, cloudlet and mobile edge computing," *2017 Global Internet of Things Summit (GIoTS).*, Geneva, 2017, pp. 1–6.

[5] Tiago Gama Rodrigues, Katsuya Suto, Hiroki Nishiyama, Nei Kato, and Katsuhiro Temma, "Cloudlets activation scheme for scalable mobile edge computing with transmission power control and virtual machine migration," *IEEE Trans. Comput.*, vol. 67, no. 9, pp. 1287–1300, Sept. 2018.

[6] T. Koketsu Rodrigues, K. Suto, and N. Kato, "Edge cloud server deployment with transmission power control through machine learning for 6G Internet of Things," *IEEE Trans. Emerging Top. Comput.*, pp. 1–1, Dec. 2019.

[7] K. Tange, M. De Donno, X. Fafoutis and N. Dragoni, "A systematic survey of industrial Internet of Things security: Requirements and fog computing opportunities," *IEEE Commun. Surv. Tutorials.*, vol. 22, no. 4, pp. 2489–2520, Jul. 2020.

[8] M. Mukherjee, L. Shu and D. Wang, "Survey of fog computing: Fundamental, network applications, and research challenges," *IEEE Commun. Surv. Tutorials*, vol. 20, no. 3, pp. 1826–1857, Mar. 2018.

[9] C. Mouradian, D. Naboulsi, S. Yangui, R. H. Glitho, M. J. Morrow and P. A. Polakos, "A comprehensive survey on fog computing: State-of-the-art and research challenges," *IEEE Commun. Surv. Tutorials*, vol. 20, no. 1, pp. 416–464, Jan.–Mar. 2018.

[10] Z. Zhao, S. Bu, T. Zhao, Z. Yin, and M. Peng, "On the design of computation offloading in fog radio access networks," *IEEE Veh. Technol. Mag.*, vol. 68, no. 7, pp. 7136–7149, Jul. 2019.

[11] Y. Shi, J. Zhang, K. B. Letaief, B. Bai, and W. Chen, "Large-scale convex optimization for ultra-dense cloud-RAN," *IEEE Wireless Commun.*, vol. 22, no. 3, pp. 84–91, Jun. 2015.

[12] H. Zhang, Y. Qiu, X. Chu, K. Long, and V. C. M. Leung, "Fog radio access networks: Mobility management, interference mitigation and resource optimization," *IEEE Wireless Commun.*, vol. 24, no. 6, pp. 120-127, Dec. 2017.

[13] S. Park, O. Simeone, and S. Shamai Shitz, "Joint optimization of cloud and edge processing for fog radio access networks," *IEEE Trans. Wireless Commun.*, vol. 15, no. 11, pp. 7621–7632, Nov. 2016.

[14] J. Liu, B. Bai, J. Zhang, and K. B. Letaief, "Cache placement in Fog-RANs: From centralized to distributed algorithms," *IEEE Trans. Wireless Commun.*, vol. 16, no. 11, pp. 7039–7051, Nov. 2017.

[15] Y. Zhou, F. R. Yu, J. Chen, and Y. Kuo, "Communications caching and computing for next generation HetNets," *IEEE Trans. Wireless Commun.*, vol. 25, no. 4, pp. 104–111, Aug. 2018.

[16] B. Bai, W. Li, L. Wang, and G. Zhang, "Coded caching in Fog-RAN: *b*-matching approach," *IEEE Trans. Commun.*, vol. 67, no. 5, pp. 3753–3767, May 2019.

[17] Y. Ye, M. Xiao and M. Skoglund, "Mobility-aware content preference learning in decentralized caching networks," *IEEE Trans. Cognit. Commun. Networking*, vol. 6, no. 1, pp. 62–73, Mar. 2020.

[18] Z. M. Fadlullah and N. Kato, "HCP: Heterogeneous computing platform for federated learning based collaborative content caching towards 6G networks," *IEEE Trans. Emerging Top. Comput.*, pp. 1–1, Aprl. 2020.

[19] H. Xiang, S. Yan and M. Peng, "A realization of fog-RAN slicing via deep reinforcement learning," *IEEE Trans. Wireless Commun.*, vol. 19, no. 4, pp. 2515–2527, Apr. 2020.

[20] S. F. Abedin, M. G. R. Alam, S. M. A. Kazmi, N. H. Tran, D. Niyato, and C. S. Hong, "Resource allocation for ultra-reliable and enhanced mobile broadband IoT applications in fog network," *IEEE Trans. Commun.*, vol. 67, no. 1, pp. 489–502, Jan. 2019.

[21] L. Qi, M. Peng, Y. Liu and S. Yan, "Advanced user association in non-orthogonal multiple access-based fog radio access networks," *IEEE Trans. Wireless Commun.*, vol. 67, no. 12, pp. 8408–8421, Dec. 2019.

[22] Y. Zhou, F. R. Yu, J. Chen and Y. Kuo, "Resource allocation for information-centric virtualized heterogeneous networks with in-network caching and mobile edge computing," *IEEE Trans. Veh. Technol.*, vol. 66, no. 12, pp. 11339–11351, Dec. 2017.

[23] C. Liang, F. R. Yu, H. Yao and Z. Han, "Virtual resource allocation in information-centric wireless networks with virtualization," *IEEE Trans. Veh. Technol.*, vol. 65, no. 12, pp. 9902–9914, Dec. 2016.

[24] Y. Tang, P. Yang, W. Wu, J. W. Mark and X. Shen, "Interference mitigation via cross-tier cooperation in heterogeneous cloud radio access networks," *IEEE Trans. Cognit. Commun. Networking*, vol. 6, no. 1, pp. 201–213, Mar. 2020.

[25] H. Zhang, X. Liu, K. Long, A. Nallanathan, and V. C. M. Leung, "Energy efficient resource allocation and caching in fog radio access networks," *2018 IEEE Global Communications Conference (GLOBECOM)*, pp. 1–6, 2018.

[26] X. Huang, W. Fan, Q. Chen and J. Zhang, "Energy-efficient resource allocation in fog computing networks with the candidate mechanism," *IEEE Internet Things J.*, vol. 7, no. 9, pp. 8502–8512, Sept. 2020.

[27] J. Li, H. Chen, Y. Chen, Z. Lin, B. Vucetic and L. Hanzo, "Pricing and resource allocation via game theory for a small-cell video caching system," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 8, pp. 2115–2129, Aug. 2016.

[28] M. Hajimirsadeghi, N. B. Mandayam and A. Reznik, "Joint caching and pricing strategies for popular content in information centric networks," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 3, pp. 654–667, Mar. 2017.

[29] H. Zhang, M. Feng, K. Long, G. K. Karagiannidis, V. C. M. Leung, and V. Poor, "Energy efficient resource management in SWIPT enabled heterogeneous networks with NOMA," *IEEE Trans. Wireless Commun.*, vol. 19, no. 2, pp. 835–845, Feb. 2020.

[30] Z. Huang, R. Hu, Y. Guo, and E. Chan-Tin, "DP-ADMM: ADMM-Based distributed learning with differential privacy," *IEEE Trans. Inf. Forensics Secur.*, vol. 15, no. 8, pp. 1002–1012, Jul. 2019.

[31] S. Boyd et al, "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Found. Trends Mach. Learn.*, vol. 1, no. 3, pp. 1–122, 2011.

[32] S. Boyd, "Convex Optimization," *Cambridge University Press*, 2004.