IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received May 9, 2020, accepted June 3, 2020, date of publication June 8, 2020, date of current version June 22, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3000832

Joint Allocation on Communication and Computing Resources for Fog Radio Access Networks

YINGTENG MA, HAIJUN WANG, JUN XIONG, (Member, IEEE), JIETAO DIAO, AND DONGTANG MA, (Senior Member, IEEE)
College of Electronic Science and Technology, National University of Defense Technology, Changsha 410073, China

Corresponding authors: Jietao Diao (911903072@qq.com) and Dongtang Ma (dongtangma@nudt.edu.cn)

This work was supported by the National Natural Science Foundation of China under Grant 61372099, 61931020 and 61601480.

ABSTRACT To improve the user experience, an increasing number of mobile applications offload their computing tasks to servers with powerful computing capabilities. The fog radio access network (F-RAN) incorporates the concept of "fog computing" into the access network architecture, endowing an edge network with computing, storage, communication and control functions. In this paper, we consider a multiple fog access point (F-AP) and a multiuser F-RAN, where each user generates two different tasks: communication and computation. To satisfy the diverse quality of service requirements of different users, we jointly optimize the spectrum access, computation offloading and radio resource allocation. The problem is modeled as a mixed integer nonlinear programming problem, which is difficult to solve. In view of this, we propose a genetic algorithm based on convex optimization, i.e., the genetic convex optimization algorithm (GCOA), which divides the mixed integer nonlinear programming problem into two parts, i.e., optimization and convex optimization, to solve it in polynomial time. Simulation results are provided to verify the effectiveness of the algorithm.

INDEX TERMS Fog radio access network, resource allocation, access strategy, offload selection, genetic convex optimization algorithm.

I. INTRODUCTION

In the fifth generation (5G) mobile radio system, the rapid popularity of mobile terminals will greatly challenge the existing communication infrastructure and network topology [1]. At the same time, the development of diverse Internet mobile applications, such as the Internet of Things (IoT), social networks, and real-time video communications, places higher requirements on the transmission rate and network capacity. Some researchers have stated that from 2015 to 2020, the expected network traffic will increase by more than 1000 times [2], [3]. Cloud radio access network (C-RAN) has the characteristics of high capacity, low latency, high energy efficiency and flexible deployment. C-RAN has become one of the solutions for future 5G cellular networks that can provide high energy efficiency and high data rate [4]–[6].

The fog radio access network (F-RAN) is an extension of the edge of the cloud radio access network. It refers to a large number of processors with computing, storage, communication and other functions, called the fog-access point (F-AP, also edge computing node), placed at the edge of a cloud network (e.g., cell base stations and routers), forming a heterogeneous network in which cloud servers, F-APs and users coexist. It is considered to be data intensive and delay sensitive for a large number of mobile users [7]–[16].

In recent years, researchers from academia and industry have studied a wide range of issues related to the F-RAN. They may include system and network modeling, optimal control, multiuser resource allocation, implementation, and standardization. In [17], [18], and [19], the joint allocation of radio and computing resources to save time and energy when offloading tasks to the cloud was considered. Similarly, the authors of [20] studied joint resource allocation

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

and collaborative offloading methods in the F-RAN. Separable tasks can be performed cooperatively in the fog and cloud, but the authors did not explain how to determine the cooperative factors between fog and cloud networks. In [21], the authors studied the problem of delay-aware wireless communication resource allocation and used Lyapunov's method to design a resource allocation scheme that can reduce system delay. The authors in [22] proposed a computing resource allocation algorithm in a multi-F-AP scenario. This algorithm can overcome the limitation of the computing capacity of a single F-AP and obtain a delay performance gain. In addition, [23] proposed a wireless communication resource allocation algorithm that can meet the constraints of system computing resources and designed a loosely coupled architecture to reduce the burden on the fronthaul link and to achieve ultralow latency. References [24] and [25] studied the optimization of wireless and computing resources to minimize energy consumption in a single-cell and multicell network under a given delay constraint by deleting them jointly. However, all tasks in this case are performed in the cloud, and the access relationship between the user and the base station is set in advance. References [26] and [27] considered the fog node cooperation mode and selected the appropriate number of fog nodes to perform user computing tasks on the premise of meeting the communication resource constraints. References [28] and [29] considered the optimization problem of minimizing the sum of the energy and delay consumed by offloading in the case of multiple users, a fog node and a cloud server. Since only one fog node exists, there is no need to consider optimizing user and fog node access. Reference [30] is similar to [28] and [29] but considers users with multiple fog nodes, where user tasks are not further transmitted to the cloud and fog processing scenarios. In [31], the research on access scheduling and interference coordination methods in heterogeneous cellular networks was conducted.On this basis, some studies have proposed a layered cloud computing system. The layered cloud computing system can use both cloud computing and fog computing resources. Despite providing users with flexible choices, due to the distinctive features of different offload strategies, optimizing offload decisions is complicated. In [32], radio resource management was optimized to maximize its own quality of experience (QoE), and a distributed algorithm was designed for fixed task allocation scenarios. In [33]–[38], the joint design of resource allocation and offload determination was studied. In particular, an algorithm based on the branch and bound concept was proposed in [33] to seek the best solution for the offload cost. Based on the simplified radio resource model, [34] and [35] combined computing and communication resources to optimize the computing offload decision and designed a suboptimal algorithm to reduce the computing complexity. A similar algorithm was studied in [36] to ensure fairness. In [37], a distributed optimization framework was designed based on the alternating direction method of multipliers (ADMM), and in [38], it was extended to a nonorthogonal multiple access scheme.

To the best of our knowledge, the types of tasks considered in the above work are computing tasks, that is, communication resources exist to satisfy the demands of computing tasks. Even if the communication resources are allocated, it is also necessary to better meet the needs of the computing tasks, such as obtaining lower latency or lower energy consumption. There are few articles that consider the types of tasks that are mainly based on communication needs (such as high-speed audio and video calls). Few articles have considered how communication tasks and computing tasks coexist in the same system and how to balance the resources used by these two types of tasks to meet the user QoS. In addition, the resource constraints considered in the above work are mostly computing resources, communication resources, computing offload, energy consumption, etc.

This work studied cloud wireless access systems with cloud servers, multiple users and multiple fog nodes. Each user has to share the tasks performed by the fog wireless access network and interacts with only one of the many fog nodes. Depending on the computing power of the fog node and the load of the fronthaul link, tasks can be performed in the fog node or the cloud, and users compete with each other. By adjusting the user association and computing offload and combining communication resource allocation and computing resource allocation to achieve a balance between communication load and computing load, the maximum communication rate is achieved, while the computing delay is minimized. Through resource allocation and the balance between the communication load and computing load, the maximum communication rate and the minimum computing delay can be achieved.

The contributions of this article include the following aspects:

- Motivated by improving the quality of service of users, this paper jointly considers communication and computing resource allocation in the F-RAN. We jointly consider edge computing, cloud computing, edge computing task migration, and network spectrum resource allocation to establish the system model. By adjusting the user association and computing offload, and by combining communication resource and computing resource allocation to achieve a balance between the communication load and computing load, the objective is to maximize the user rate of all communication tasks and minimize the delay of all users to generate computing tasks.
- We formulate a mixed integer nonlinear programming (MINLP) problem. Considering the 0-1 integer constraints of user association, computing migration, computing resources, and frequency band resource allocation in the model, drawing on the idea of the branch and bound method, a genetic convex optimization algorithm for hierarchical optimization of user association and resource allocation is proposed.
- We carry out simulation analysis in the case of computing tasks and communication tasks of different priorities

and analyze the different impacts of F-AP nodes, radio frequency remote head(RRH) nodes, bandwidth, computing resources and task importance on resource allocation.

- We numerically prove that the algorithm proposed herein has polynomial-level computational complexity and that its performance is better than that of discrete particle swarm optimization, the greedy heuristic algorithm and the traditional genetic algorithm.

The rest of this paper is organized as follows. Section II introduces the system model. We formulate the MINLP problem in section III. We propose a genetic convex optimization algorithm in section IV. Numerical results are provided in section V. Finally, we conclude the paper in section VI. A list of notation and abbreviations used throughout the paper is provided in Table 1.

II. SYSTEM MODEL

The system model is shown in Figure 1. We consider a multi-cell solution with F-RAN architecture. This solution consists of the following network equipment: M F-APs, R RRHs, 1 HPN, and K randomly distributed F-UEs. F-AP set, RRH set and F-UE set are expressed as $\mathbb{M} = \{1, 2, \cdots, m, \cdots M\}$, $\mathbb{R} = \{1, 2, \cdots, r, \cdots R\}$ and $\mathbb{E} = \{1, 2, \cdots, e, \cdots E\}$, respectively. There are N RF remote heads, where $N = M + R$. The entire available spectrum bandwidth is B Hz.

Radio frequency remote head, which mainly includes RF module, related amplifier/filter and antenna. It also includes digital signal processing, digital/analog conversion, analog-to-digital conversion and other modules. The radius of the cell is denoted as d_{BS}. The computing power of each F-AP is the same, and the computing rate is f_{F-AP}. F-AP and RRH have the same service radius d_{RRH}.

F-UE i randomly generates the following two tasks, which are represented by a task indicator $e_i \in \{0, 1\}$, where $e_i = 0$ indicates that the F-UE i requests a computing task, and $e_i = 1$ indicates that the F-UE i requests a communication task. Computing tasks and communication tasks are modeled as follows.

A. COMPUTING TASKS

In order to simplify the model, it is assumed that the offloading strategy of the F-UE itself is known, and the computing tasks retained in the F-UE itself are not considered. All the computing tasks expressed are tasks that need to be offloaded to the F-AP or BBU pool. Computational tasks are expressed using a hard deadline task model. Expressed as follows, the computing task can be represented by a three-field symbol $c = \{L, \tau_D, X\}$.This common symbol contains information on the task input data size L (in bits), completion deadline τ_D (in seconds), and computing intensity X (CPU cycles per bit). It is required that task must be completed before the hard deadline τ_D.

TABLE 1. Summary of notations in this paper.

Notation	Description
MINLP	mixed integer nonlinear programming
F-RAN	fog radio access networks
F-AP	fog-access point
F-UE	fog user equipment
RRH	Radio frequency remote head
HPN	high power node
QoS	quality of service
GCOA	convex optimization-based genetic algorithm
GA	genetic algorithm
GHA	greedy heuristic algorithm
DPSO	discrete particle swarm optimization algorithm
$\mathbb{M}$	F-AP set
$\mathbb{R}$	RRH set
$\mathbb{E}$	F-UE set
d_{BS}	the radius of the cell
f_{F-AP}	the computing rate
d_{RRH}	service radius of F-AP and RRH
e_i	Task indicator, user i generates a computing task when equal to 1, and generates a communication task when equal to 0
L	the task input data size
τ_D	completion deadline
X	computing intensity
$C_i^{\min}$	the user's minimum transmission rate
L^*	user association indication matrix
L^+	the association indication matrix for F-UE i to be connected to the BBU pool through the RRH and F-AP
$c_{i,n}$,$c_{i,H}$	Spectrum efficiency when FUE i is connected to RF module
$h_{i,n}$,$h_{i,H}$	Channel gain when FUE i is connected to RF module
$p_{i,n}$,$p_{i,H}$	Transmit power when FUE i is connected to RF module
$\widetilde{h_{i,n}}$,$\widetilde{h_{i,H}}$	Rayleigh random variables
α	path loss constant
ξ	the log-normal shadow
$G_{i,n}$,$G_{i,H}$	antenna gains
$d_{i,n}$	the distance between the user and the RRH
B_i	denote the bandwidth allocated to user i
R_i	transmission rate from F-UE i to F-AP
$R_{i,n}$	transmission rate from F-UE i to F-AP n
$R_{i,H}$	transmission rate from F-UE i to F-AP n
$a_{i,m}$	proportion of computing resources allocated by F-AP m to F-UE i
$\tau_{i,m}$	computing delay caused by offloading when computing tasks are offloaded to F-AP
$\tau_{i,m}^u$	computing delay caused by communication when computing tasks are offloaded to F-AP
τ_i^a	total delay when computing tasks are offloaded to F-AP
$\tau_{i,n}^H$	delay from F-AP to BBU pool
τ_i^b	total delay when computing tasks are offloaded to the BBU pool through F-AP(RRH)
τ_i^c	total delay when computing tasks are directly offloaded to the BBU pool

There are three ways to complete a computing task. The first is to execute in the F-AP, the second is to offload the computing tasks to the BBU pool for execution through F-AP or RRH, and the third is to directly offload the computing tasks to the BBU pool for execution through HPN.

B. COMMUNICATION TASKS

In actual scenarios, there are many tasks that require little or no computing resources. For example, when a user needs to perform audio and video communication, the main task requirement is the bandwidth requirement, that is, stable and

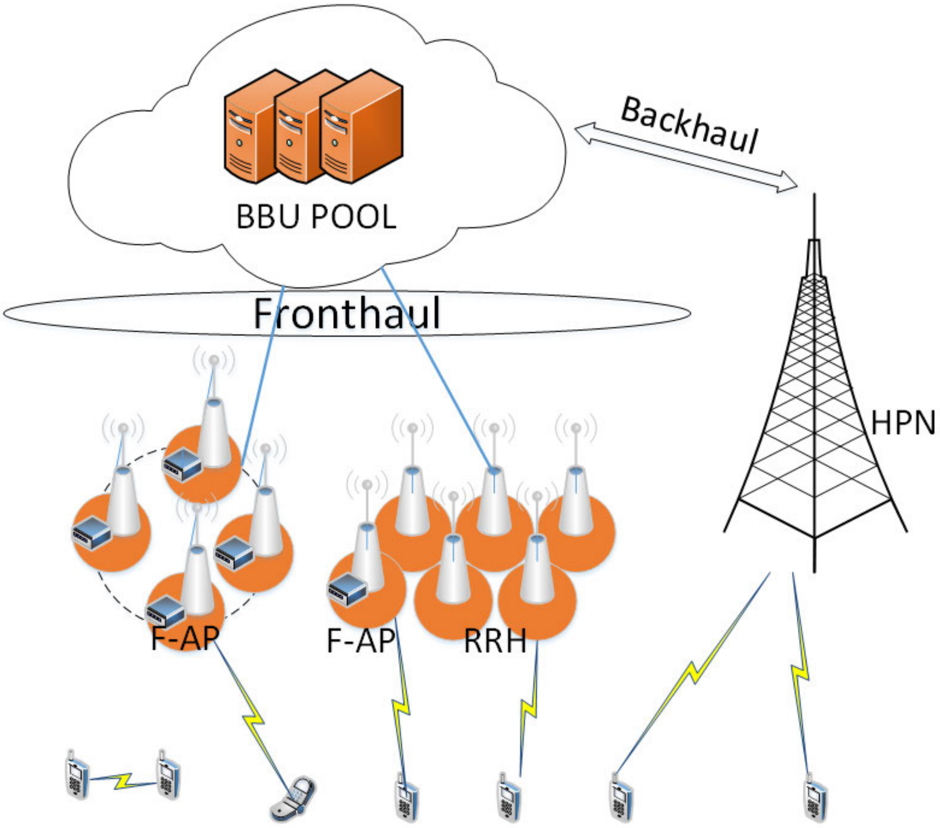


FIGURE 1. System model.

sufficient bandwidth is required for data transmission. This creates a communication task with communication resources as the main constraint. Users can use their own computing power to meet the computing resource requirements without occupying the computing resources in the F-AP and BBU pools in the system.The communication task of user i in this paper can be expressed by the user's minimum transmission rate $C_i^{\min}$.

This paper balanced the QoS of two different tasks by controlling the user association, allocated bandwidth, computing resources, and uploading policies.

III. PROBLEM FORMULATION

Denote $e_i \in \{0, 1\}$ as the task indicator, where $e_i = 0$ indicates that F-UE i generates a communication task, and where $e_i = 1$ indicates that F-UE i generates a computing task.

A. COMMUNICATION MODEL

The F-UE selects different association methods according to actual requirements. It is assumed here that the user is a single antenna user and can only connect to one RRH or F-AP. In addition, users who do not belong to any RRH or F-AP can also directly connect with HPN. Therefore, the user association indication matrix can be defined as

$$L^* = \begin{bmatrix} l_{11} & \cdots & l_{1M} & l_{1(M+1)} & \cdots & l_{1(M+R)} & l_1 \\ l_{21} & \cdots & l_{2M} & l_{2(M+1)} & \cdots & l_{2(M+R)} & l_2 \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \vdots \\ l_{E1} & \cdots & l_{EM} & l_{E(M+1)} & \cdots & l_{E(M+R)} & l_E \end{bmatrix} \tag{1}$$

where $l_{ij} \in \{0, 1\}$, $l_{im} = 1$ indicates that the user is connected to the $F - AP_m$, and $l_{i(M+e)} = 1$ indicates that the user i is connected to the RRH_e and $l_i = 1$ indicates that user i is directly connected to HPN. And only one value of each row of the matrix is 1, and the rest are all *0*.

The association matrix between users, RRH and F-APs is

$$L^{-} = \begin{bmatrix} l_{11} & \cdots & l_{1M} & l_{1(M+1)} & \cdots & l_{1(M+R)} \\ l_{21} & \cdots & l_{2M} & l_{2(M+1)} & \cdots & l_{2(M+R)} \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ l_{E1} & \cdots & l_{EM} & l_{E(M+1)} & \cdots & l_{E(M+R)} \end{bmatrix} \tag{2}$$

In addition, if the computing resources need to be offloaded to the BBU pool, the corresponding RRH or F-AP must also be connected to the BBU pool. Use $L^+ = \left[l_1^E \; l_1^E \; \cdots \; l_E^E\right]^T$ to indicate the association indication matrix for F-UE i to be connected to the BBU pool through the RRH and F-AP. $l_i = 1$ to indicate that the user needs to connect to the BBU pool through the RRH or F-AP, other-wise it is not required.

The spectral efficiency of F-UE i connected to RF unit n and the spectral efficiency of F-UE directly connected to HPN can be expressed as

$$c_{i,n} = \log_2\left(1 + \frac{p_{i,n}h_{i,n}}{N_0}\right) \tag{3}$$

$$c_{i,H} = \log_2\left(1 + \frac{p_{i,H}h_{i,H}}{N_0}\right) \tag{4}$$

where $p_{i,n}$ is the transmission power when the F-UE is connected to the RRH, $p_{i,H}$ is the transmission power when the F-UE is connected to the HPN, $h_{i,n}$ is the channel gain between the i-th cell and the n-th radio frequency unit, and $g_{i,H}$ is the F-UE and The channel gain of the HPN connection can be expressed as

$$h_{i,n} = \widetilde{h_{i,n}}\xi G_{i,n}\left(\frac{d_0}{d_{i,n}}\right)^{\alpha} \tag{5}$$

$$h_{i,H} = \widetilde{h_{i,H}}\xi G_{i,H}\left(\frac{d_0}{d_{i,H}}\right)^{\alpha} \tag{6}$$

where $\widetilde{h_{i,n}}$ and $\widetilde{h_{i,H}}$ are Rayleigh random variables, α is the path loss constant, ξ is the log-normal shadow, antenna gains are $G_{i,n}$ and $G_{i,H}$, and $d_{i,n}$ is the distance between the user and the RRH.

The frequency band resources are divided and allocated to the end users as needed, and the end users access the RRHs and F-AP in the same manner as FDMA. Let B_i denote the bandwidth allocated to user i, then the transmission rate from F-UE i to F-AP is

$$R_i = \sum_n l_{i,n}R_{i,n} \tag{7}$$

where $R_{i,n} = c_{i,n}B_i$.

The transmission rate of F-UE i directly to the BBU pool is

$$R_{i,H} = l_i c_{i,H} B_i \tag{8}$$

Fibers are used to connect from the F-AP to the BBU pool, and the rate assigned to each task is fixed at R_e.

B. COMPUTING MODEL

Computing tasks can be performed in F-AP or BBU pools, and the most important consideration for computing tasks is to complete the task within a hard deadline.

1) IF F-UE COMPUTING TASKS ARE OFFLOADED TO F-AP

Let f_{F-AP} denote the total computing resources of F-AP i, and $a_{i,m} \in [0, 1]$ denote the proportion of computing resources allocated by F-AP m to F-UE i. The computing delay caused by offloading the task to the F-AP is

$$\tau_{i,m} = \frac{X}{f_{F-AP}a_{i,m}} \tag{9}$$

The amount of data uploaded by the F-UE i for computing is L, then the delay caused by communication is

$$\tau_{i,m}^{u} = \frac{L}{R_{i,m}} \tag{10}$$

The total delay is

$$\tau_i^a = \sum_m l_{i,m}\left(\tau_{i,m} + \tau_{i,m}^u\right) \tag{11}$$

2) F-UE COMPUTING TASKS ARE OFFLOADED TO THE BBU POOL THROUGH F-AP(RRH)

Assume that the computing capacity in the BBU pool is ideal, that is, the computing capacity is sufficient. If the computing tasks are offloaded into the BBU pool, the computing delay is approximately zero. The delay consumption is mainly concentrated on the communication requirements.

In this case, the tasks to be transmitted not only need to be transmitted to the F-AP(RRH), but also need to be relayed to the BBU pool through the F-AP(RRH). The amount of data that F-UE i needs to upload is calculated as: L, then the delay from F-AP to BBU pool is:

$$\tau_{i,n}^{H} = \frac{L}{R_e} \tag{12}$$

The total delay is:

$$\tau_i^b = \sum_n l_{i,n}\left(\tau_{i,n}^u + l_i^H\tau_{i,n}^H\right) \tag{13}$$

3) F-UE COMPUTING TASKS ARE DIRECTLY OFFLOADED TO THE BBU POOL

The delay loss only includes the transmission delay, which is:

$$\tau_i^c = l_i\frac{L}{R_{i,H}} \tag{14}$$

So, no matter how the tasks are related and where they are performed, the total delay is

$$\tau_i = \sum_n l_{i,n}\tau_{i,n}^u + l_i^H\tau_{i,n}^H + \left(1 - l_i^H\right)\tau_{i,n} + \tau_i^c \tag{15}$$

In order to minimize the computing delay of the computing task and maximize the total rate of communication tasks, this chapter considers user association, computation migration, spectrum, and computational resource allocation modeling together. The problems formed are as P0.

$$\begin{aligned}
P0:\quad & \max_{l,B,a}\lambda\sum\nolimits_i\left(1-e_i\right)\left(R_i + R_{i,H}\right) - (1-\lambda)\sum\nolimits_i e_i\tau_i \\
s.t.\quad & \\
C1:\quad & R_i > C_i^{\min},\quad \forall i \\
C2:\quad & \sum\nolimits_i a_{i,m} \leqslant 1,\quad \forall n \\
C3:\quad & \tau_i \leqslant \tau_i^{\max} \\
C4:\quad & \sum\nolimits_i B_i \leqslant B \\
C5:\quad & l \in \{0, 1\} \\
C6:\quad & 0 \leqslant a_{i,m} \leqslant e_i \\
C7:\quad & 0 \leqslant a_{i,m} \leqslant l_{i,n} \\
C8:\quad & 0 \leqslant l_{i,n} \leqslant \frac{d_{RRH}}{d_{i,n}} \\
C9:\quad & 0 \leqslant l_{i,n}^H \leqslant e_i
\end{aligned} \tag{16}$$

where, e_i is a task indicator, and λ is the importance of the communication task. C1 indicates that the communication rate of all users who generate communication tasks must be greater than the required minimum value $C_i^{\min}$. C2 represents a computing resource constraints of F-AP. C3 indicates that the maximum delay of the user who generates the computing task cannot exceed the rated delay $\tau_i^{\max}$. C4 represents the constraint of total radio resources. C5 uses the *0-1* variable to indicate the user's access policy. C6 and C7 represent constraints of available computing resources allocated to user *i*. C8 indicates that when the service distance d_{RRH} is greater than the distance $d_{i,n}$ between the user and the RRH, the user and the RRH can establish a connection. C9 indicates that the offload indicator will be *1* only when a computing task is generated.

Since the backpack problem is a complete NP-hard problem, and the problem P0 is an extension of the backpack problem, the problem P0 is an NP-hard problem.

IV. PROPOSED ALGORITHM

In this section, through analyzing and transforming the optimization problem, a convex optimization-based genetic algorithm (GCOA) is proposed.

By observing the problem P0, we know that P0 is a MINLP problem which is NP-hard. A common solution is to use the branch and bound method to get an acceptable approximate solution. However, the computational complexity of the branch and bound method is only reduced by *2-3* times compared to the brute force search algorithm, and it is still exponential. With the increase in the number of users and F-APs, the computational complexity is unacceptable. Therefore, it can only be solved using the greedy algorithm. At present, the genetic algorithm and particle swarm algorithm perform better in this respect.

First convert the problem P0 and expand it to get in (17), as shown at the bottom of this page.

By observing P0, it can be known that among all unknown constraint variables, l is an integer variable of 0-1, while spectrum resource B and computing resource a are continuous variables, and integer variables are coupled with each other and between integer variables and continuous variables. The module is a MINLP problem. The general solution to the MINLP problem is the branch and bound(BB) method. In the worst case, the workload required by BB increases exponentially as the problem size increases. Therefore, when the number of users and the number of base stations increase, BB cannot be an effective solution.

When the correlation matrix l is determined, the problem P0 can be converted into the problem P1

$$\begin{aligned} P1: \quad & \max_{B,a} \lambda \sum\nolimits_i (1-e_i)\left[\left(\sum\nolimits_n c_{i,n}B_i + c_{i,E}B_i\right)\right] - (1-\lambda) \\ & \cdot \sum\nolimits_i e_i \left[\sum_n \left[\frac{L}{c_{i,n}B_i} + \frac{L}{R_e} + \frac{X}{f_{F-AP}a_{i,n}} + l_i \frac{L}{c_{i,H}B_i}\right]\right] \\ s.t. \quad & C1-C4, C6, C7 \end{aligned} \tag{18}$$

The inverse function in form is a convex function. Since the non-negative weighted sum and composition with affine mapping are operations that preserve the convexity of the function, all constraints of the problem P1 can be equivalently converted to some form where the convex function is less than or equal to a constant. Obviously, the feasible region of the problem P1 is a convex set, and the objective function of the problem P1 is a convex function. Therefore, the problem P1 is a convex optimization problem, which can be solved by a recent convex optimization algorithm.

In this way, the problem can be divided into two parts: solving the convex optimization of each association matrix and finding the optimal association matrix. Genetic algorithm can be used to find the optimal correlation matrix, and each correlation matrix is solved by convex optimization method. Formation of genetic convex optimization algorithms (GCOA).

The detailed steps are as follows.

A. CHROMOSOME EXPRESSIONS

The main goal of the genetic algorithm is to find the optimal power correlation matrix when the optimal power allocation scheme for each connection mode is known. Therefore, the user association matrix L^* and the upload indication matrix L^+ can be used as chromosome expressions, i.e.,

$$L^* = \begin{bmatrix} l_{11} & \cdots & l_{1M} & l_{1(M+1)} & \cdots & l_{1(M+R)} & l_1 \\ l_{21} & \cdots & l_{2M} & l_{2(M+1)} & \cdots & l_{2(M+R)} & l_2 \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \vdots \\ l_{E1} & \cdots & l_{EM} & l_{E(M+1)} & \cdots & l_{E(M+R)} & l_E \end{bmatrix} \tag{19}$$

where $l_{ij} \in \{0, 1\}$, $l_{i,m} = 1$ indicate that the user is connected to the $F-AP_m$, and $l_{i,M+e} = 1$ indicates that the user i is connected to the RRH_e and $l_i = 1$ indicates that user i is directly connected to HPN. And only one value of each row of the matrix is *1*, and the rest are all *0*. $L^+ = \left[l_1^E \; l_2^E \; \cdots \; l_E^E\right]^T$ indicates the association indication matrix that F-UE i needs to connect to the BBU pool through the F-AP, and $l_i = 1$ indicates that the user needs to connect to the BBU pool through the F-AP, otherwise it is not needed.

$$\begin{aligned} P0: \quad & \max_{l,B,a} \lambda \sum\nolimits_i (1-e_i)\left[\left(\sum\nolimits_n \left(l_{i,n}c_{i,n}\right)B_i + l_i c_{i,E} B_i\right)\right] - (1-\lambda)\sum\nolimits_i e_i \left[\sum\nolimits_n \left[\begin{matrix} l_{i,n}\dfrac{L}{c_{i,n}B_i} + l_i^H \dfrac{L}{R_e} + \left(1-l_i^H\right) \\ \cdot \dfrac{X}{f_{F-AP}a_{i,n}} + l_i \dfrac{L}{c_{i,H}B_i} \end{matrix}\right]\right] \\ s.t. \quad & C1-C9 \end{aligned} \tag{17}$$

B. SELECTION OF PRIMARY POPULATIONS

The selection of the initial population will affect the convergence speed and convergence point. In this paper, all users are directly connected to the BBU pool, all users access according to optimal channel conditions, and all computing tasks are offloaded to the BBU pool for computing, and all users access and calculate according to optimal channel conditions The tasks are all calculated in the F-AP node as three primary chromosomes, and the rest are randomly generated.

C. GENETIC MANIPULATION

Genetic operations are a crucial step in genetic algorithms. Divided into retention and cross mutation.

In this paper, the top 25% of the optimal chromosomes are retained, but only the optimal one is fully retained. The remaining superior chromosomes contain excellent genes, so they are not completely replaced. Instead, it randomly changes the access base station of one of the users, or changes the computing offloading strategy of the F-AP. The specific change strategy of the computing offloading will be described in detail in the subsequent mutation.

1) **Crossover:** According to the roulette selection method, chromosomes are selected for cross mutation, and the probability of each chromosome being selected is $p_i = \frac{ad_i}{\sum ad}$. Randomly select a column in the middle of the chromosome as the intersection point. A gene fragment of a chromosome with a higher fitness function is retained before the intersection, and a gene fragment of a chromosome with a lower fitness function is retained after the intersection to obtain a new chromosome.
2) **Variation:** Variation of chromosomes is an important means to ensure the diversity of chromosomes. In this design, there are two factors that affect the final result, one is the association strategy between the user and the F-AP (RRH), and the other is the F-AP computing offload strategy. Set the mutation number seed $rand1 \in \{0, 1, 2\}$. If $rand1 = 0$, no mutation operation is performed. If $rand1 = 1$, the correlation matrix mutation operation is performed. If $rand1 = 2$, the mutation operation of the upload matrix is performed. The correlation matrix mutation operation is as follows. Set the mutation number seed $rand2 \in \{0, 1\}$. If $rand2 = 0$, randomly change the value of a position of the correlation matrix and observe the value of the upload parameter at that position. If the upload parameter is *0*, it will be maintained. If the upload parameter is *1*, the value of the upload parameter is randomly taken. If $rand2 = 1$, the value of the association matrix is changed randomly. The upload matrix mutation operation is as follows. Set the mutation number seed $rand3 \in \{0, 1, 2\}$. If $rand3 = 0$, invert all values of the upload matrix. If $rand3 = 1$, randomly change the upload matrix value. If $rand3 = 2$, then upload all Set the matrix value to *0*.

D. EVALUATION FUNCTION

The design of the evaluation function will affect the results of the algorithm. It can be seen from the observation that the objective function of this design is the problem of maximum value. The higher the value, the better the effect, so it can be used directly as the evaluation function in this article. However, it is known through experiments that directly using the evaluation function will cause the bandwidth to be concentrated on users and base stations with good channel conditions. The bandwidth allocated by other users can only meet their own needs. In practical applications, we always hope that the user experience of each user is not much different. In addition, because the communication rate is generally in Mbps, and the value of the computing requirement often exists between 0.1-0.2s, in order to prevent the final result due to the magnitude of different task values, the objective function needs to be redesigned. The value of the objective function of this design is as follows:

$$ad = 0.1 \times \log\left[\lambda \sum_i (1 - e_i)\left(R_i + R_{i,H}\right)\right] - (1 - \lambda)\sum_i e_i \tau_i \tag{20}$$

The reason for the design is as follows: Through computing, we find that with the support of the same bandwidth, the value of the communication rate is $C_i \in [1, 100]$ MBps, and the value of the delay is calculated as $\tau_i \in [0.1, 0.2]$ s. By taking the logarithmic method, the communication rate is almost equal to the calculated rate value, and the bandwidth is not allocated to a certain user as much as possible, resulting in a large user experience gap.

Algorithm 1 shows the process of the proposed algorithm.

Algorithm 1 Genetic Convex Optimization Algorithm

```
Initialization:
Get the values required by the algorithm: B, R, F, $e_i$, X, L, $c_{i,n}$, $c_{i,E}$
Generation of primary genetic factors
Calculate the optimal bandwidth allocation strategy $B_i^0$ and optimal computing resource allocation strategy $a_i^0$ using convex optimization according to each initial genetic factor
Calculate the fitness function $ad_i^0$ of each primary genetic factor
Set the number of iterations n
while $m < n$ do
    Calculate the probability of selecting each genetic factor $P_i$ based on $ad_i^{m-1}$
    Pick the top 25% of the best genetic factors
    Keep the best fit
    The remaining 25% randomly change the base station accessed by a user or change the F-AP computing offload strategy
    Generate Variation Seed $rand1 \in \{0, 1\}$
    if $rand1 = 1$ then
```

```
14:       Randomly change the base station access strategy of a user
15:     else
16:       Change the computing offloading strategy of F-AP
17:       Generate upload mutation random seed $rand3 \in \{0, 1, 2\}$
18:       if $rand3 = 0$ then
19:         $l_i^E = 1 - l_i^E$
20:       else
21:         if $rand3 = 1$ then
22:           Randomly changing an $l_i^E$
23:         else
24:           $l_i^E = 0$
25:         end if
26:       end if
27:     end if
28:     Genetic factors that cross over based on $P_i$ selection
29:     Crossing selected genetic factors pairwise
30:     Generate Variant Random Integer $randint \in \{0, 1, 2\}$
31:     if $randint = 0$ then
32:       Does not change
33:     else
34:       if $randint = 1$ then
35:         Mutation Association Matrix
36:         Generate Variant Random Number Seed $rand2 \in \{0, 1\}$
37:         if $rand2 = 0$ then
38:           Randomly change the association strategy of a user
39:         else
40:           Change Association Policy for All Users
41:         end if
42:       else
43:         Change the computing offloading strategy of F-AP, like 18-24
44:       end if
45:     end if
46:     Best bandwidth allocation strategy $B_i^m$ and best computing resource allocation strategy $a_i^m$
47:     Calculate the fitness function $ad_i^m$ for each new genetic factor
48:   end while
49:   output: $L^*, L^+, B_i, a_i$
50:   Calculate the target value
```

V. SIMULATION RESULTS

This section analyzes the performance of the proposed algorithm based on simulation results. Consider a typical F-RAN network with an area size of $1000m1000m$. The network contains *1* macro base station, *2* F-APs, *2* RRHs, and several users randomly distributed in the base station area. Each F-AP is equipped with an edge computing server, which can provide computing services for user-generated computing tasks. RRH does not have an edge computing server, and can only handle the communication tasks of forwarding users. All small base stations are connected to the BBU pool using optical fiber, and the optical fiber allocates a fixed bandwidth for each task. The computing power of the BBU pool is assumed to be ideally infinite. Assume that each user randomly generates one of two tasks, a computing task and a communication task. All generated communication tasks require a minimum rate of 1MBps or more. The size of the data required for the generated computing tasks is randomly distributed between [100, 1000] KB, and the number of CPU computing cycles required for the task size is distributed between [0.2, 1] Gcycles. The transmission power of each user is 100mW, and the maximum tolerance delay for the completion of computing tasks performed by the terminal is randomly distributed between 0.2 s and 0.1 s. The main parameters used in the simulation are shown in Table 2.

TABLE 2. Units for magnetic properties.

Parametersl	Reference Value
Total system bandwidth	50MHz
User transmit power	100mW
Background noise power spectral density	-169dBm/Hz
F-AP computing rate	50GHz
Forward rate	20MBps

A. PERFORMANCE ANALYSIS UNDER DIFFERENT NUMBER OF USERS AND F-APS NUMBERS

Figures 2 and 3 respectively show the impact of different numbers of F-APs on computing tasks and communication tasks when the pay the same attention to communication tasks and computing tasks. When the number of RRHs increases to four and the number of F-APs decreases to zero, the system degenerates into the H-CRAN model.

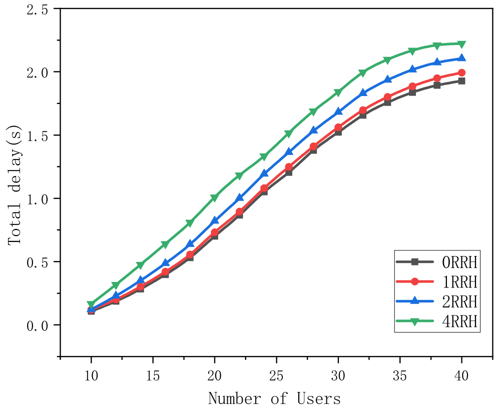


FIGURE 2. The total computing delay of the system.

Figure 2 shows the total computing delay of the system as the number of user terminals increases. As can be seen from the Figure 2, as the number of users increases, the total delay in computing increases regardless of the number of F-APs. With the same number of users, as the number of RRHs increases, the number of F-APs decreases, and the total system delay increases continuously. When the number of terminals is 10, the number of F-APs has little effect on the computing of the total delay. This is because the total number of computing tasks is small and the corresponding amount of computing resources is relatively small. As the number of terminals increases, the total amount of computing resources required by the system continues to increase. The more F-APs, the less time required for computing. However, the total computing delay difference reaches the maximum when the number of users is 25, and when the number of users is ⩾ 25, the increase in the computing delay difference caused by the increase in user terminals under different F-AP numbers is almost zero. This is because when the number of user terminals increases, the utilization of local computing resources has reached its maximum.

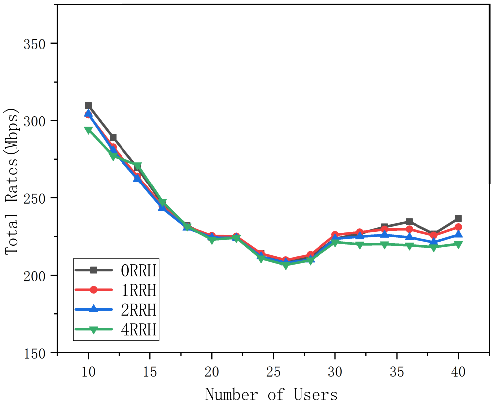


FIGURE 3. The total communication rate of the system.

Figure 3 shows how the total communication rate of the system changes as the number of user terminals increases. It can be seen that under the same communication and computing tasks, an increase in the number of F-APs cannot bring about a large increase in communication rate. However, the overall number of F-APs is still larger, and the overall communication rate is slightly improved. This is because the increase of computing resources makes the system not need enough bandwidth to reduce the delay of the computing task at the communication rate, and allocates more bandwidth to the communication task. In addition, it can be found that even if the number of users increases, the total communication rate of the system is still decreasing. This is because the increase in the number of users also brings an increase in computing tasks, requiring more bandwidth and communication resources to meet the low latency This reduces the bandwidth allocated to communication tasks and reduces the overall communication rate.

In order to further explore the relationship between the increase in the number of F-APs and the computing delay. In Figure 4, we study the change of the total computing delay with the increase of the number of user terminals when all computing tasks are generated in the system. It can be clearly found that the larger the number of RRHs and the smaller the number of F-APs, the greater the total system delay. However, the increase in the number of users has little effect on the total delay difference. This is because the delay reduction that F-AP computing resources can bring is limited. To further reduce the delay, you can only increase the number of F-APs or increase the computing capacity of F-APs.

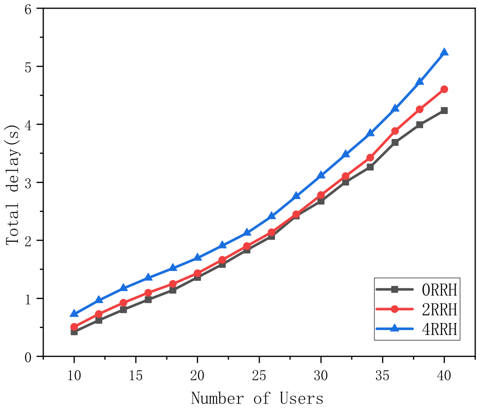


FIGURE 4. The total system delay with full generation of computing tasks.

B. IMPACT OF DIFFERENT MISSION IMPORTANCE ON SYSTEM EFFICIENCY

The degree of attention paid to different tasks will have a great impact on the allocation of resources in the system, which will affect the total computing delay and communication rate of the system. The situation where the two kinds of tasks have the same degree of attention has been analyzed above, and will not be repeated here.

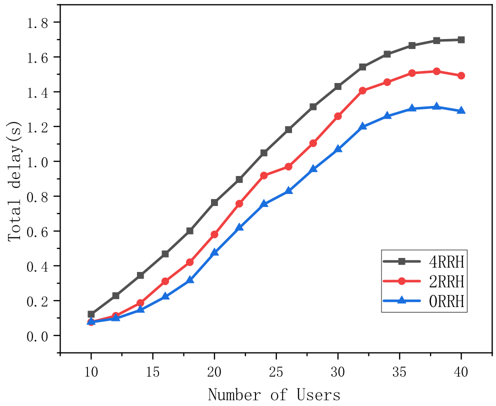


FIGURE 5. Total system delay when the importance of a computing task is much greater than the importance of a communication task.

Figures 5 and 6 respectively show the changes in the computing delay and communication rate with the increase of the number of user terminals when, that is, the attention of computing tasks is much greater than the attention of communication tasks. And that means, compared to the increase in bandwidth of communication users, it is more desirable for computing users to obtain lower latency.

Figure 5 shows the total computing delay of the system as the number of user terminals increases. It can be clearly seen that when the number of RRHs is larger and the number of F-APs is smaller, the total system delay is increasing.

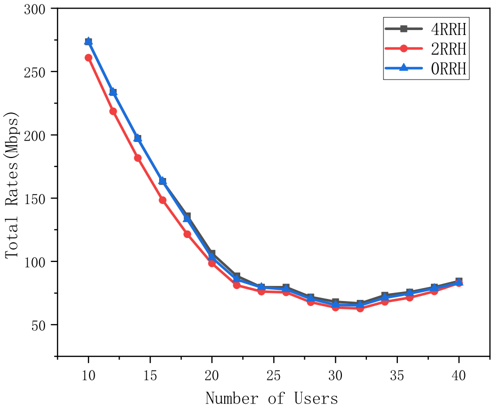


FIGURE 6. Total system rate when the importance of a computing task is much greater than the importance of a communication task.

And this phenomenon is gradually obvious as the number of user terminals increases.

Figure 6 shows how the total communication rate of the system changes as the number of user terminals increases. It can be seen that when the attention of computing tasks is much greater than the attention of communication tasks, the increase of the F-AP ratio in the system has no gain in the increase of the communication rate.

It can be seen that when the attention of computing tasks is much greater than the attention of communication tasks, the increase in the proportion of F-AP in the system greatly improves the system gain, and the computing can be greatly improved when the communication rates are not much different Capacity, thereby shortening the time delay, so you can use as many F-AP nodes as possible when biasing towards computational tasks.

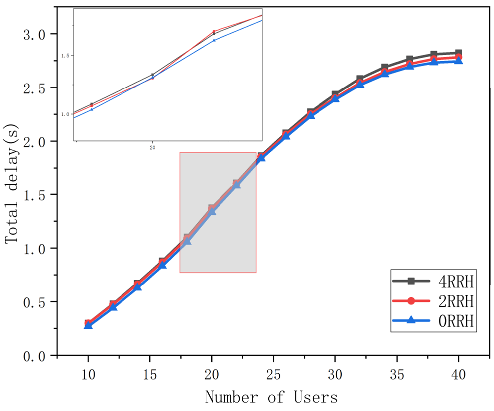


FIGURE 7. Total system delay when the importance of a computing task is much less than the importance of a communication task.

Figures 7 and 8 respectively show the changes of the computing delay and communication rate with the increase of the number of user terminals when, that is, the attention of computing tasks is much less than the attention of communication tasks. And that means, compared to computing users to obtain lower latency, it is more desirable for the increase in bandwidth of communication users

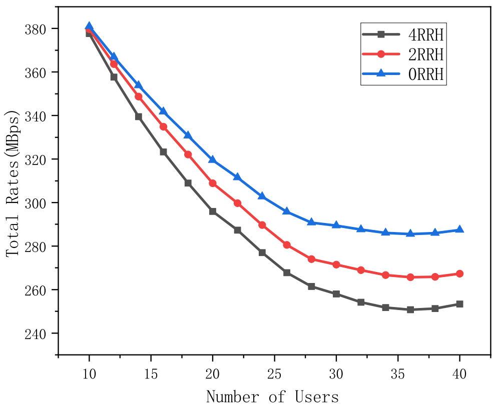


FIGURE 8. Total system rate when the importance of a computing task is much less than the importance of a communication task.

Figure 7 shows the change of the total computing delay of the system as the number of user terminals increases. It can be seen that when the attention of computing tasks is much less than the attention of communication tasks, the bandwidth allocated to computing tasks only meets the minimum requirements of computing tasks. Therefore, regardless of the number of F-AP nodes, the total computing delay is almost equal, depending on Calculate the delay threshold of a task.

Figure 8 shows how the total communication of the system changes as the number of user terminals increases. It can be seen that even though we pay more attention to the communication tasks, as the number of users increases, the system with more F-APs has a larger total communication rate. This is because due to the increase in computing power, the delay caused by computing is smaller. Under the same delay requirement, less bandwidth can be allocated for computing tasks. In this way, more bandwidth can be allocated to communication tasks, thereby greatly improving the communication rate of the system.

To better understand the impact of different attention levels on system performance between tasks, Figures 9 and 10 use four F-AP nodes uniformly, and simulates the performance of the system with attention levels on system performance between tasks Changes in the number of users. $lamuda = \frac{\lambda}{1-\lambda}$, The larger the lambda, the greater the attention paid to communication tasks.

Figure 9 shows how the total communication rate of the system changes with the number of users under different attention levels situations. It can be seen that, As expected, the higher the focus on communication tasks, the higher the overall communication rate. As the number of users increases, the difference in overall communication rates becomes more apparent. However, it is worth noting that the increase in the attention of the communication task results in a smaller increase in the total communication rate compared to the decrease in the total communication rate due to the increase in the attention of the computing task.

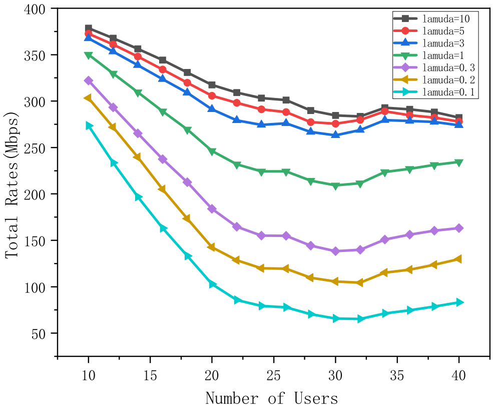


FIGURE 9. Total communication rate of the system changes with the number of users under different mission importance situations.

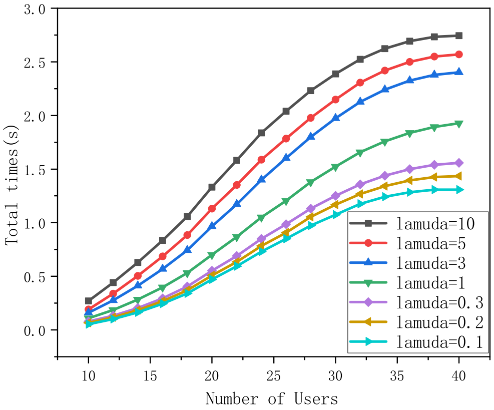


FIGURE 10. Total system delay of the system changes with the number of users under different mission importance situations.

Figure 10 shows how the total system latency varies with the number of users for different task concerns. It can be clearly seen that the higher the attention of the computing task, the smaller the total system delay. And as the number of users increases, the difference in total latency becomes more apparent. But again, the delay caused by the increased attention of the communication task is increased compared with the decrease of the delay caused by the increased attention of the computing task.

C. THE IMPACT OF THE IMPROVEMENT OF THE FRONTHAUL LINK RATE ON THE SYSTEM EFFICIENCY

The fronthaul rate will have a greater impact on system performance. Figures 11 and 12 simulate the changes in the total information rate and total delay of the system with the increase of the fronthaul rate when there are 4 F-APs.

Figure 11 shows how the total system delay varies with user growth when the fronthaul link rate increases from 5MBps to 50MBps. It can be clearly seen that the higher the fronthaul rate, the lower the total system delay. This is because the increase in the fronthaul link rate has brought about a reduction in the time cost of offloading computing tasks from the F-AP to the BBU pool. As a result, the delay in offloading

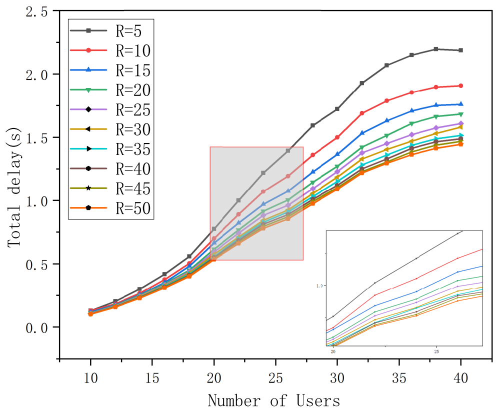


FIGURE 11. The total system delay varies with user growth when the fronthaul link rate increases from 5MBps to 50MBps.

the F-AP node to the BBU pool including users is reduced. As more tasks are offloaded to the BBU pool for execution, each F-AP can allocate more computing resources to tasks left for local execution. This results in time and cost savings.

Figure 12 shows the change of the total information rate of the system with the increase of users under different fronthaul rates. It can be seen that the increase in the fronthaul rate will bring a certain increase in the overall system rate, but the increase is not significant.

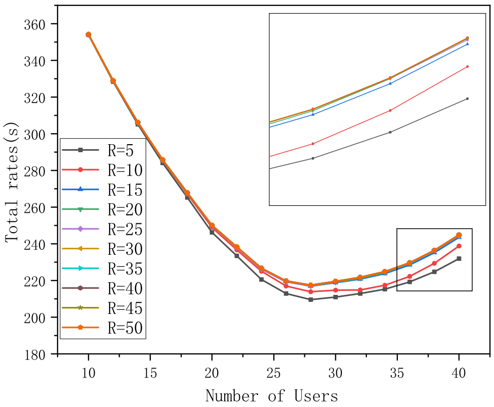


FIGURE 12. The total communication rate varies with user growth when the fronthaul link rate increases from 5MBps to 50MBps.

D. THE IMPACT OF DIFFERENT RESOURCE ALLOCATION STRATEGIES ON SYSTEM PERFORMANCE

This section compared the proposed algorithm with the user access methods for users accessing according to the channel optimal strategy.

In the simulation in this section, there are four F-AP nodes, excluding RRH nodes, the forward rate R = 20, and the task importance index lamuda = 1.

Figure 13 shows how the total system delay varies with the number of user terminals under different allocation strategies. It can be seen that when the number of users is small, the user accesses on the optimal channel, and the total delay of the

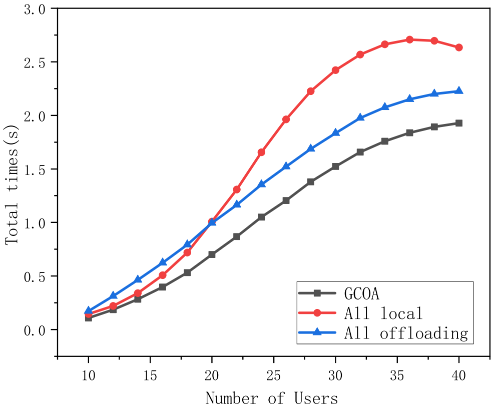


FIGURE 13. The total system delay varies with the number of user terminals under different allocation strategies.

system where all computing tasks are performed locally is lower than the full offload to the BBU pool for execution on the optimal channel access. of. This is because the computing resources in the F-AP node are sufficient, so the delay caused by computing is lower than the delay caused by offloading tasks to the BBU pool. This situation starts to change when the number of users is N = 20. This is because the computational load of the F-AP has reached its maximum. The computing delay caused by continuing to offload tasks to the F-AP node is greater than the delay caused by communication offloading it to the BBU. However, the algorithm proposed in this paper is better than the two methods proposed above in the case of large or small number of users.

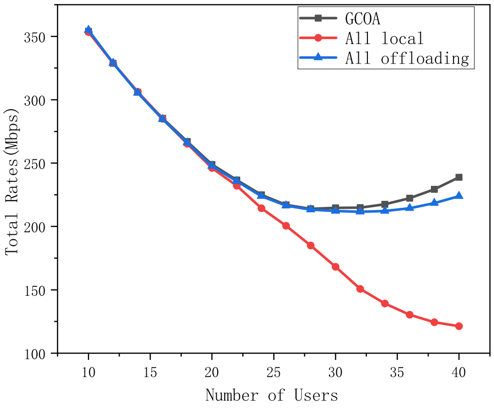


FIGURE 14. The total communication rate varies with the number of user terminals under different allocation strategies.

Figure 14 shows the change of the total system rate with the increase of the number of user terminals in the case of different allocation strategies. It can be seen that when the number of users is small, the communication rates of the three offload situations are not much different. This is because all users access according to the optimal channel situation, and as the number of users increases, the three The gap has gradually widened. Among them, the algorithm proposed in this paper performs best. This is because with the increase in the number of users, the access strategy depends not only on the channel gain, but also takes into account load balancing, and offloads some computing tasks to suboptimal channel conditions, but the computing tasks assume fewer nodes. As a result, the computing delay is reduced, the bandwidth allocated to the computing task is saved, and more bandwidth is allocated to the communication task.

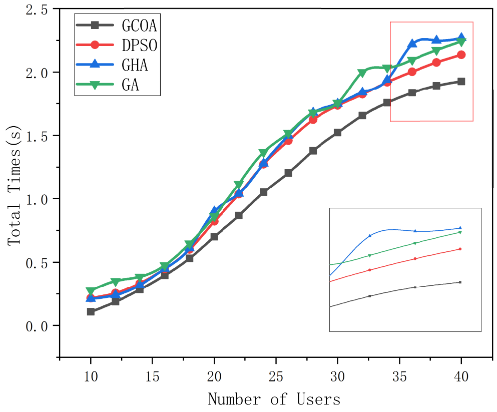


FIGURE 15. The total system delay varies with the number of user terminals for different algorithm.

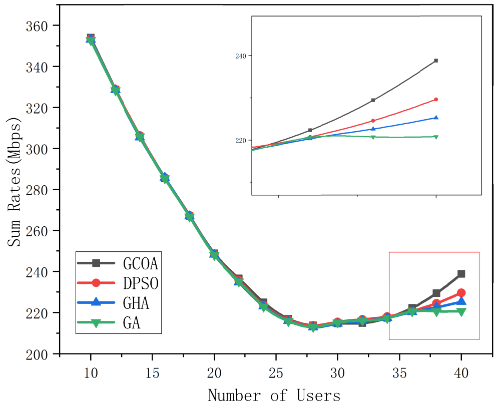


FIGURE 16. The Sum Rates varies with the number of user terminals for different algorithm.

E. COMPARISON BETWEEN THE ALGORITHM PROPOSED IN THIS PAPER AND DISCRETE PARTICLE SWARM OPTIMIZATION

This section compared the proposed algorithm with other commonly used optimization algorithms. For the MINLP problem formed by computing unloading, the currently commonly used algorithm is the greedy heuristic algorithm(GHA). In addition, the discrete particle swarm optimization algorithm(DPSO) and the genetic algorithm(GA) have a good performance on this problem. In the simulation in this section, there are 4 F-AP nodes, excluding RRH nodes, the forward rate R = 20, and the task importance index lamuda = 1.

Figure 15 shows how the total delay of the deffirents algorithms changes with the increase in the number of user terminals. It can be seen that as the number of users increases, the total delay obtained by the algorithm proposed in this paper is significantly better than that of other algorithms.

Figure 16 shows the changes in the total system rate of different algorithms as the number of user terminals increases. It can be seen that as the number of users increases, the total system rate using the algorithm proposed in this article gradually outperforms other algorithms.

VI. CONCLUSION

In this paper, we considered the joint resource allocation of computing and communication in F-RAN. In order to satisfy the QoS of users of different types of tasks, we formulated the MINLP problem with the limited processing power and fronthaul of each F-AP. A genetic convex optimization algorithm considering user access, computing offload, computing resource allocation, and spectrum resource allocation is proposed to obtain a feasible suboptimal solution. The simulation analyzes the impact of different number of users, different F-AP numbers, two different kinds of tasks with different attention levels, different fronthaul link rates, and different resource allocation strategies on system performance. The simulation results showed that the genetic convex optimization algorithm has a better payoff than than discrete particle swarm optimization, greedy heuristic algorithm and traditional genetic algorithm.

REFERENCES

[1] M. Armbrust, R. G. A. Fox, A. D. Joseph, R. H. Katz, A. Konwinski, G. Lee, D. A. Patterson, A. Rabkin, I. Stoica, and M. Zaharia. (Feb. 2012). *Above the Clouds: A Berkeley View of Cloud Computing*. [Online]. Available: https://www2.eecs.berkeley.edu/Pubs/TechRpts/ 2009/EECS-2009-28.eps

[2] Q. Zhang, L. Cheng, and R. Boutaba, "Cloud computing: State-of-the-art and research challenges," *J. Internet Services Appl.*, vol. 1, no. 1, pp. 7–18, May 2010.

[3] M. Chiang and T. Zhang, "Fog and IoT: An overview of research opportunities," *IEEE Internet Things J.*, vol. 3, no. 6, pp. 854–864, Dec. 2016.

[4] M. Peng, Y. Li, J. Jiang, J. Li, and C. Wang, "Heterogeneous cloud radio access networks: A new perspective for enhancing spectral and energy efficiencies," *IEEE Wireless Commun.*, vol. 21, no. 6, pp. 126–135, Dec. 2014.

[5] Y. Zhang and Y. Wang, "A framework for energy efficient control in heterogeneous cloud radio access networks," in *Proc. IEEE/CIC Int. Conf. Commun. China (ICCC Workshops)*, Jul. 2016, pp. 1–5.

[6] Z. Yu, K. Wang, H. Ji, X. Li, and H. Zhang, "Dynamic resource allocation in TDD-based heterogeneous cloud radio access networks," *China Commun.*, vol. 13, no. 6, pp. 1–11, Jun. 2016.

[7] M. Patel *et al.*, "Mobile-edge computing—Introductory technical white paper," ETSI, Sophia Antipolis, France, White Paper, Sep. 2014.

[8] G. P. Fettweis, "The tactile Internet: Applications and challenges," *IEEE Veh. Technol. Mag.*, vol. 9, no. 1, pp. 64–70, Mar. 2014.

[9] A. Al-Fuqaha, M. Guizani, M. Mohammadi, M. Aledhari, and M. Ayyash, "Internet of Things: A survey on enabling technologies, protocols, and applications," *IEEE Commun. Surveys Tuts.*, vol. 17, no. 4, pp. 2347–2376, 4th Quart., 2015.

[10] Juniper. (Mar. 2015). *Smart Wireless Devices and the Internet of Me*. [Online]. Available: http://itersnews.com/wp-content/uploads/experts/2015/03/96079Smart-Wireless-Devices-and-the-Internet-of-Me.eps

[11] Y. Bonan, A. Yuan, and P. Mugen, "Fog computing based radio access networks: Architecture, principles and challenges," *Telecommun. Sci.*, vol. 32, no. 6, pp. 20–27, 2016.

[12] Y. C. Hu, M. Patel, D. Sabella, N. Sprecher, and V. Young, *Mobile Edge Computing a Key Technology Towards 5G*, 1st ed. 2015.

[13] W. Shi and S. Dustdar, "The promise of edge computing," *Computer*, vol. 49, no. 5, pp. 78–81, May 2016.

[14] O. Salman, I. Elhajj, A. Kayssi, and A. Chehab, "Edge computing enabling the Internet of Things," in *Proc. IEEE 2nd World Forum Internet Things (WF-IoT)*, Milan, Italy, Dec. 2015, pp. 603–608.

[15] A. Ahmed and E. Ahmed, "A survey on mobile edge computing," in *Proc. 10th Int. Conf. Intell. Syst. Control (ISCO)*, Coimbatore, India, Jan. 2016, pp. 1–8.

[16] T. Taleb, K. Samdanis, B. Mada, H. Flinck, S. Dutta, and D. Sabella, "On multi-access edge computing: A survey of the emerging 5G network edge cloud architecture and orchestration," *IEEE Commun. Surveys Tuts.*, vol. 19, no. 3, pp. 1657–1681, 3rd Quart., 2017.

[17] H. Liu, F. Eldarrat, H. Alqahtani, A. Reznik, X. de Foy, and Y. Zhang, "Mobile edge cloud system: Architectures, challenges, and approaches," *IEEE Syst. J.*, vol. 12, no. 3, pp. 2495–2508, Sep. 2018.

[18] M. T. Beck, M. Werner, S. Feld, and S. Schimper, "Mobile edge computing: A taxonomy," in *Proc. Int. Conf. Adv. Future Internet (AFIN)*, Lisbon, Portugal, Nov. 2014, pp. 48–54.

[19] P. Mach and Z. Becvar, "Mobile edge computing: A survey on architecture and computation offloading," *IEEE Commun. Surveys Tuts.*, vol. 19, no. 3, pp. 1628–1656, 3rd Quart., 2017.

[20] D. Sabella, A. Vaillant, P. Kuure, U. Rauschenbach, and F. Giust, "Mobile-edge computing architecture: The role of MEC in the Internet of Things," *IEEE Consum. Electron. Mag.*, vol. 5, no. 4, pp. 84–91, Oct. 2016.

[21] E. Ahmed and M. H. Rehmani, "Mobile edge computing: Opportunities, solutions, and challenges," *Future Gener. Comput. Syst.*, vol. 70, pp. 59–63, May 2017.

[22] T. X. Tran, A. Hajisami, P. Pandey, and D. Pompili, "Collaborative mobile edge computing in 5G networks: New paradigms, scenarios, and challenges," *IEEE Commun. Mag.*, vol. 55, no. 4, pp. 54–61, Apr. 2017.

[23] S. Wang, X. Zhang, Y. Zhang, L. Wang, J. Yang, and W. Wang, "A survey on mobile edge networks: Convergence of computing, caching and communications," *IEEE Access*, vol. 5, pp. 6757–6779, Jun. 2017.

[24] M. Salmani and T. N. Davidson, "Multiple access computational offloading with computation constraints," in *Proc. IEEE Workshop Signal Process. Adv. Wireless Commun.*, Jul. 2017, pp. 385–389.

[25] S. Sardellitti, G. Scutari, and S. Barbarossa, "Joint optimization of radio and computational resources for multicell mobile-edge computing," *IEEE Trans. Signal Inf. Process. Over Netw.*, vol. 1, no. 2, pp. 89–103, Jun. 2015.

[26] T.-C. Chiu, W.-H. Chung, A.-C. Pang, Y.-J. Yu, and P.-H. Yen, "Ultra-low latency service provision in 5G fog-radio access networks," in *Proc. IEEE 27th Annu. Int. Symp. Pers., Indoor, Mobile Radio Commun. (PIMRC)*, Sep. 2016, pp. 1–6.

[27] A.-C. Pang, W.-H. Chung, T.-C. Chiu, and J. Zhang, "Latency-driven cooperative task computing in multi-user fog-radio access networks," in *Proc. IEEE 37th Int. Conf. Distrib. Comput. Syst. (ICDCS)*, Jun. 2017, pp. 615–624.

[28] M.-H. Chen, B. Liang, and M. Dong, "A semidefinite relaxation approach to mobile cloud offloading with computing access point," in *Proc. IEEE 16th Int. Workshop Signal Process. Adv. Wireless Commun. (SPAWC)*, Jun. 2015, pp. 186–190.

[29] M.-H. Chen, M. Dong, and B. Liang, "Joint offloading decision and resource allocation for mobile cloud with computing access point," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2016, pp. 3516–3520.

[30] T. Quang Dinh, J. Tang, Q. Duy La, and T. Q. S. Quek, "Offloading in mobile edge computing: Task allocation and computational frequency scaling," *IEEE Trans. Commun.*, vol. 65, no. 8, pp. 3571–3584, Aug. 2017.

[31] L. Zhou, X. Hu, E. C.-H. Ngai, H. Zhao, S. Wang, J. Wei, and V. C. M. Leung, "A dynamic graph-based scheduling and interference coordination approach in heterogeneous cellular networks," *IEEE Trans. Veh. Technol.*, vol. 65, no. 5, pp. 3735–3748, May 2016.

[32] H. Shah-Mansouri and V. W. S. Wong, "Hierarchical fog-cloud computing for IoT systems: A computation offloading game," *IEEE Internet Things J.*, vol. 5, no. 4, pp. 3246–3257, Aug. 2018.

[33] E. El Haber, T. M. Nguyen, and C. Assi, "Joint optimization of computational cost and devices energy for task offloading in multi-tier edge-clouds," *IEEE Trans. Commun.*, vol. 67, no. 5, pp. 3407–3421, May 2019. [Online]. Available: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8626532

[34] M.-H. Chen, M. Dong, and B. Liang, "Resource sharing of a computing access point for multi-user mobile cloud offloading with delay constraints," *IEEE Trans. Mobile Comput.*, vol. 17, no. 12, pp. 2868–2881, Dec. 2018.

[35] M.-H. Chen, B. Liang, and M. Dong, "Multi-user multi-task offloading and resource allocation in mobile cloud systems," *IEEE Trans. Wireless Commun.*, vol. 17, no. 10, pp. 6790–6805, Oct. 2018.

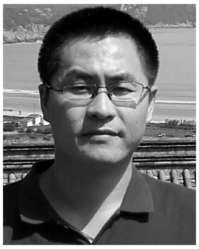
[36] J. Du, L. Zhao, J. Feng, and X. Chu, "Computation offloading and resource allocation in mixed fog/cloud computing systems with min-max fairness guarantee," *IEEE Trans. Commun.*, vol. 66, no. 4, pp. 1594–1608, Apr. 2018.

[37] Y. Wang, X. Tao, X. Zhang, P. Zhang, and Y. T. Hou, "Cooperative task offloading in three-tier mobile computing networks: An ADMM framework," *IEEE Trans. Veh. Technol.*, vol. 68, no. 3, pp. 2763–2776, Mar. 2019. [Online]. Available:https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8607120

[38] Y. Liu, F. R. Yu, X. Li, H. Ji, and V. C. M. Leung, "Distributed resource allocation and computation offloading in fog and cloud networks with non-orthogonal multiple access," *IEEE Trans. Veh. Technol.*, vol. 67, no. 12, pp. 12137–12151, Dec. 2018.

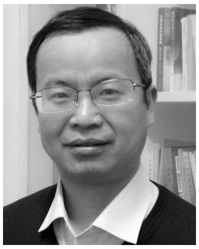
YINGTENG MA received the B.S. degree from the College of Electronic Science and Technology, National University of Defense Technology, Changsha, China, in 2015, where he is currently pursuing the M.S. degree. His research interests include cognitive radio networks and fog radio access networks.

HAIJUN WANG received the B.S. degree in Internet of Things engineering from Shandong University, in 2014, and the M.S. degree in information and communication engineering from the National University of Defense Technology, Changsha, China, in 2016, where he is currently pursuing the Ph.D. degree with the College of Electronic Science and Technology. His research interests include cognitive radio networks and unmanned aerial vehicle communications.



JUN XIONG (Member, IEEE) received the B.S. and Ph.D. degrees from the National University of Defense Technology (NUDT), China, in 2009 and 2014, respectively. He is currently an Associate Professor with the College of Electronic Science and Technology, NUDT. His research interests include cooperative communications, physical layer security, and resource allocation, where he has published more than 40 refereed articles.

JIETAO DIAO received the master's degree in information and communication engineering from the National University of Defense Technology. From 1996 to 2000, he was a Lecturer with the School of Electronic Science and Engineering, National University of Defense Technology. From 2000 to 2013, he was an Associate Professor with the School of Electronic Science and Engineering, National University of Defense Technology. In 2013, he became a Professor at the School of Electronic Science and Engineering, National University of Defense Technology. He presided over more than ten national natural science funds, 973 pre-research projects, and school pre-research projects. He was authorized two national invention patents. He has published more than 20 high-level scientific research articles, of which three were retrieved by SCI and 12 by EI.



DONGTANG MA (Senior Member, IEEE) received the B.S. degree in applied physics and the M.S. and Ph.D. degrees in information and communication engineering from the National University of Defense Technology (NUDT), Changsha, China, in 1990, 1997, and 2004, respectively. From 2004 to 2009, he was an Associate Professor with the College of Electronic Science and Engineering, NUDT. Since 2009, he has been a Professor with the Department of Cognitive Communication, School of Electronic Science and Engineering, NUDT. From August 2012 to February 2013, he was a Visiting Professor with the University of Surrey, U.K. He has published more than 150 journals and conference papers. He is one of the Executive Directors of Hunan Electronic Institute. His research interests include wireless communication and networks, physical layer security, and intelligent communication and networks. He has severed as the TPC member for PIMRC, from 2012 to 2019.