Received December 1, 2017, accepted January 1, 2018, date of publication January 8, 2018, date of current version February 28, 2018.

Digital Object Identifier 10.1109/ACCESS.2018.2790963

Fair Resource Allocation for System Throughput Maximization in Mobile Edge Computing

ZHENGFA ZHU, JUN PENG, (Member, IEEE), XIN GU, HENG LI, (Member, IEEE), KAIYANG LIU, ZHUOFU ZHOU, AND WEIRONG LIU, (Member, IEEE)
School of Information Science and Engineering, Central South University, Changsha 410000, China

Corresponding author: Weirong Liu (frat@csu.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 61379111, Grant 61772558, Grant 61672537, and Grant 61672539.

ABSTRACT Communication resource allocation is important for improving the performance of users in mobile edge computing (MEC) scenarios. In existing studies, the users in the MEC system typically suffer from unfair resource allocation, which results in the inefficient resource utilization and degraded user performance. To address this challenge, in this paper we propose a fair resource allocation approach to maximize the overall network throughput, under the constraint of each mobile user's minimum transmission rate. We formulate the problem as a fair Nash bargaining resource allocation game, and the existence and uniqueness of the solution to this game model are analyzed. By adopting the time-sharing variable, we obtain the near optimal bargaining resource allocation strategy for the mixed integer nonlinear programming optimization. The user's priority is further considered in the iterative implementation of the proposed algorithm by considering the time delay constraint of users. Simulation results show that the proposed scheme outperforms existing methods in terms of resource allocation fairness and overall system throughput.

INDEX TERMS Mobile edge computing, resource allocation, fairness, system throughput maximization, minimum rate requirement.

I. INTRODUCTION

In recent years, mobile edge computing (MEC) has been considered as a promising technology to support the next generation Internet, such as Tactile Internet, Internet of Things (IoT), and Internet of Me, by migrating the mobile computing, network control and storage to the network edges [1]. Therefore, it is possible to run the highly demanding applications at the user equipments while meeting strict delay requirements [2].

With the development of mobile network and mobile devices, the number of mobile internet users has shown explosive growth, which results in the spectrum scarcity for mobile users [3]. The spectrum scarcity has raised the necessity of a fair spectrum allocation for mobile users. Without fair resource allocation, the minimum rate constraint of some users may not be satisfied, which leads to the degradation of the user performance [4].

Spectrum resource allocation has received considerable attentions in MEC systems [4]–[7]. Zhao *et al.* [4] proposed a quantitative study on adaptive resource allocation by designing a frequency reuse scheme to mitigate interference and maintain high spectral efficiency. In order to improve the performance in terms of higher system capacity, Singhal *et al.* [5] proposed a resource allocation scheme with differentiated QoS provisioning for cell-edge active users. Zhao *et al.* [6] proposed a greedy heuristic method to achieve the optimal resource allocation for users. Ren *et al.* [7] developed a piecewise resource allocation algorithm to allocate the communication and computation resources jointly.

However, most of these approaches are centralized allocation schemes without considering the channel diversity among users. The individual profit of each user may result in the deployment difficulty of the centralized allocation in MEC. Game theory is thus introduced to address the individual characteristics of mobile users in the spectrum allocation problem [8]–[10]. Chen *et al.* [8] proposed a game-based distributed algorithm to solve the resource allocation problem among multiple mobile devices with limited communication resource. Taking energy consumption and transmission delay into account, Ma *et al.* [9] developed a distributed algorithm to achieve joint radio and communication resource allocation with the game equilibrium. Xu *et al.* [10] proposed an enhanced adaptive video delivery scheme with joint cache and radio resource allocation in order to provide the low-latency and high quality services for mobile device users.

We note, however, that existing studies did not explicitly consider the fairness of the resource allocation among users. This implies that some users may be benefited while the other might be penalized, which leads to the deteriorative user experience and inefficient resource utilization [3]. To address this challenge, in this paper, we propose a Nash bargaining game based resource allocation method for mobile users in the MEC system. By considering the user's individual demand, the proposed Nash bargaining game based method maximizes the overall system throughput while satisfying the minimum delay requirements of users. Nash bargaining game can guarantee the fairness in resource allocation problems [3], [11]. For instance, Han *et al.* [3] used the Nash bargaining game to achieve the generalized proportional fairness based on optimal coalition pairs among users. Lee and Leung [11] realized the fair allocation of subcarrier and power in wireless mesh networks.

To further handle the channel diversity of users, we introduce the user priority in the subchannel allocation design by considering the user's time delay margin. The user priority is determined according to the time delay constraint of users. The smaller the delay requirement of the user, the higher the priority will be allocated. In the design, each user sorts all subchannels according to the channel conditions. All users are divided into two subsets. Two subsets of users are matched to the subchannels by the Hungarian method respectively, so that the rate of matched user-channel pair is greater than any other possible matching. Thus the user can send the data in accordance with the established priority. The users with much delay margin can share some subchannels to best-matching user, while meeting their own time delay requirement.

The contributions in this paper can be summarized as follows:

- The bandwidth and power joint allocation to maximize the system throughput is analyzed in MEC wireless network for resource limited mobile equipment. The individual variations of users are modeled.
- Taking the fairness of users and the global throughput into consideration, the resource optimization is formulated as Nash bargaining optimization problem. The existence and uniqueness of the solution are analyzed, and the fairness of the solution is proved.
- The user priority is introduced according to the users' time constraints. Based on user priority, the users with excessive time margin can allocate their redundant subcarriels to other users with more stringent time delay so the spectrum utilization can be improved.
- Due to the NP-hard characteristic of original bargaining optimization problem, by introducing the time-sharing variable, the discrete subchannel allocation is changed to allocation optimization of continuous time variable, and the original problem is transformed into standard convex optimization problem. The optimal collaborative negotiation resource allocation strategy is obtained by Karush-Kuhn-Tucker (KKT) conditions.

The rest of the paper is organized as follows. Section II presents the system model and problem formulation. The proposed scheme is designed in section III. In section IV, Numerical results are illustrated. The conclusion is discussed in section V.

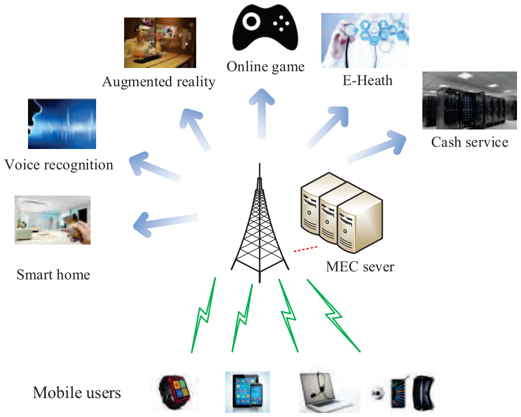


FIGURE 1. **The architecture of a mobile edge computing system.**

II. SYSTEM MODEL AND PROBLEM FORMULATION

We consider an uplink transmission scenario in mobile networks with the MEC. As shown in Fig. 1, there is one base station (BS) that works in OFDMA mode with wireless channel set $\mathcal{K} = \{1, 2, \ldots, K\}$. The set of mobile users within BS coverage area is denoted by $\mathcal{N} = \{1, 2, \ldots, N\}$. In this scenario, one subchannel can only be used by one user at a time. It is assumed that each user has a delay-sensitive computation task to be completed on the mobile device or on the mobile edge cloud server via computation offloading. The tasks include interactive gaming, high-definition image processing, face recognition, virtual reality, and so on [8], [12]. In general, each task to be processed can be described by a tuple as $J_i = \{d_i, \omega_i\}$, $i \in \mathcal{N}$, where d_i denotes the size of computation input data, including the program codes and input parameters, and ω_i denotes the total number of CPU cycles required to accomplish this task. In this paper, it is assumed that the battery-powered mobile device has sufficient energy to support task offloading or local execution [12]. Then, we discuss the computation overhead of time cost for both local execution and offloading approaches. The main notations adopted in this paper are presented in Table 1.

A. SYSTEM MODEL

For local task execution, we consider a mobile device that can handle multiple tasks simultaneously with parallel computing. In general, modern processors of mobile devices have the dynamic voltage and frequency scaling (DVFS) ability. Therefore, the processors can distribute their computing capacity to different tasks. Here, the local computing capacity

TABLE 1. Symbols and definitions.

Symbol	Definition
K	Number of subchannels
$\mathcal{K}$	Set of subchannels
N	Number of users
$\mathcal{N}$	Set of users
J_i	Task of user i
d_i	Size of computation input data for task J_i
ω_i	Total number of CPU cycles required to accomplish task J_i
f_i^l	The local computing capacity distributed to task J_i by user i
t_i^l	Computation execution time to accomplish task J_i locally
$r_{i,k}$	Uplink transmission rate of user i on subchannel k
R_i	Total uplink transmission rate of user i
$p_{i,k}$	Transmit power of user i on subchannel k
$g_{i,k}$	Channel gain of user i on subchannel k
σ_k^2	Additive white Gaussian noise power on subchannel k
f_i^c	Computing capacity for task J_i provided by MEC
t_i^c	Execution time for offloading task J_i
t_i^{tr}	Data transmission time for offloading task J_i
t_i^s	The total cost time for offloading task J_i, including the transmission time and execution time
w_0	Bandwidth per subchannel
$T_i^{\max}$	Maximum time delay constraint for user i
$P_i^{\max}$	Maximum transmission power constraint for user i

distributed to task J_i is denoted by f_i^l. Then the task local execution time t_i^l is computed as

$$t_i^l = \omega_i / f_i^l. \tag{1}$$

For the edge computing, user i can choose to offload its computation task J_i to the MEC server. We denote transmission power of user i on subchannel k as $p_{i,k}$. According to the Shanon-Hartley formula, the achievable rate $r_{i,k}$, $k \in \mathcal{K}$, on each subchannel is given by

$$r_{i,k} = w_0 \log_2 \left(1 + p_{i,k} g_{i,k} / \sigma_k^2\right), \tag{2}$$

where w_0 is the subchannel bandwidth, $g_{i,k}$ and σ_k^2 denote the channel gain and additional white Gaussian noise power on subchannel k, respectively. Let $a_{i,k} \in \{0, 1\}$ denotes whether subchannel k is assigned to user i or not (i.e., if subchannel k is assigned to user i, $a_{i,k} = 1$; otherwise, $a_{i,k} = 0$). Then, the total uplink transmission rate R_i of user i can be expressed as

$$R_i = \sum_{k=1}^{K} a_{i,k} w_0 \log_2 \left(1 + p_{i,k} g_{i,k} / \sigma_k^2\right). \tag{3}$$

And with the input data size d_i for offloading, the transmission time t_i^{tr} is calculated as

$$t_i^{tr} = d_i / R_i. \tag{4}$$

Let f_i^c be the computation capacity that the MEC server assigns to user i to execute task J_i. Then the task execution time t_i^c on the MEC server is calculated as

$$t_i^c = \omega_i / f_i^c. \tag{5}$$

In many mobile applications, the output of the computation is often of considerably small size, so the transmission delay for output feedback of J_i can be ignored [12].

According to the analysis above, the time cost t_i^s for the case where user i chooses offloading the task is given by

$$t_i^s = t_i^{tr} + t_i^c. \tag{6}$$

It is assumed that the maximum time delay each user i can accept to accomplish offloading task is $T_i^{\max}$, and the maximum time delay should be smaller than that the task executed locally on mobile user itself, i.e. $T_i^{\max} < t_i^l$. Furthermore, it must be satisfied that the time cost on computation task offloading is no more than the maximum time delay each user i can accept, which means

$$t_i^s \leq T_i^{\max}. \tag{7}$$

Moreover, in order to achieve the desired time delay constraint $T_i^{\max}$, user i supplies the maximum power $P_i^{\max}$ to accomplish task J_i during the whole time delay $T_i^{\max}$.

It is assumed that sufficient computational resource is available at the MEC server, such that any of resource requirements of executing any offloading task J_i can be satisfied. Therefore, the execution time delay t_i^c on MEC server side is small enough compared to the transmission time delay t_i^{tr}, thus t_i^c can be negligible [12]. Then, the time delay for computation task offloading equals the time delay for the input data transmission, i.e.,

$$t_i^s = t_i^{tr} = d_i / R_i. \tag{8}$$

From (7) and (8), we have

$$R_i \geq d_i / T_i^{\max}. \tag{9}$$

Let $R_i^{\min} = d_i / T_i^{\max}$, which means the minimum uplink transmission rate for user i meeting the maximum time delay constraint $T_i^{\max}$. As d_i, $T_i^{\max}$ and $R_i^{\min}$ are constants, then we have

$$R_i \geq R_i^{\min}. \tag{10}$$

In this paper, our goal is to improve the bandwidth efficiency by maximizing the overall system rate with the minimum uplink transmission rate constraint in (10). There are two problems need to be resolved: 1) how to assign subchannels among users, and 2) how the power should be allocated to corresponding subchannels for each user i under maximum power constraint $P_i^{\max}$. These problems can be considered as generalized cases of proportionally fair resource allocation. The Nash bargaining solution (NBS) is a cooperative game theory, which has been broadly applied to resolve fair resource allocation problems [13], [14]. In this section, we will briefly review the basic definitions and concepts for Nash bargaining solution at first. Then, we will give an overview on how to apply these ideas into bandwidth and power allocation problems in this paper.

B. NASH BARGAINING GAME AND PROBLEM FORMULATION

Let $\mathcal{N}$ be the set of players in the bargaining game, which is the mobile users set in this paper. Let $\mathcal{S}$ be a closed and convex subset of $\mathcal{R}^N$ to represent the set of feasible payoff allocations that they can get if the players cooperate with each other. Let $R_i^{\min}$ be the minimum payoff that the i-th player would expect, which means the minimum uplink transmission rate user i requires. Suppose $\{R_i \in \mathcal{S} | R_i \geq R_i^{\min}, \forall i \in \mathcal{N}\}$ is a nonempty bounded set. Define $\mathbf{R}^{\min} = (R_1^{\min}, \ldots, R_N^{\min})$, then the pair of $(\mathcal{S}, \mathbf{R}^{\min})$ constructs a N-player bargaining game.

We define the Pareto efficient point [15], where a player can not find another point to improve the total utility of all the players, as a selection criterion for the bargaining solutions within the feasible set $\mathcal{S}$.

Definition 1: The point $(R_1,\ldots,R_N)$ *is said to be Pareto optimal, if and only if there does not exist other allocation* $R_i^{'}$ *such that* $R_i^{'} \geq R_i, \forall i \in \mathcal{N}$, *and* $R_i^{'} > R_i, \exists i \in \mathcal{N}$, *i.e., there is no other allocation that contributes to superior performance for some players without causing inferior performance for some other players.*

There may be even an infinite number of Pareto optimal points in a bargaining game. Thus, we need further criteria to determine the best Pareto point of the game. Here we mainly focus on fairness among players. Thus the criterion of NBS fairness is chosen to solve the resource allocation problem in wireless networks with MEC. NBS provides an unique and fair Pareto optimal point under the following axioms [15].

Definition 2: $\bar{\boldsymbol{R}}$ *is said to be a NBS in* $\mathcal{S}$ *for* $\boldsymbol{R}^{\boldsymbol{min}}$, *which means,* $\bar{\boldsymbol{R}} = \Phi(\mathcal{S}, \boldsymbol{R}^{\boldsymbol{min}})$, *if the following axioms are satisfied.*

a) Individual Rationality: $\bar{R}_i = \sum_{j=1}^K \bar{r}_{i,j} \geq R_i^{min}, \forall i \in \mathcal{N}$.

b) Feasibility: $\bar{\boldsymbol{R}} \in \mathcal{S}$.

c) Pareto Optimality: For every $\hat{\boldsymbol{R}} \in \mathcal{S}$, if $\sum_{j=1}^K \hat{r}_{i,j} \geq \sum_{j=1}^K \bar{r}_{i,j}, \forall i \in \mathcal{N}$, then $\sum_{j=1}^K \hat{r}_{i,j} = \sum_{j=1}^K \bar{r}_{i,j}, \forall i \in \mathcal{N}$.

d) Independence of Irrelevant Alternative: If $\bar{\boldsymbol{R}} \in \mathcal{S}' \subset \mathcal{S}$, $\bar{\boldsymbol{R}} = \Phi(\mathcal{S}, \boldsymbol{R}^{\boldsymbol{min}})$, then $\bar{\boldsymbol{R}} = \Phi(\mathcal{S}', \boldsymbol{R}^{\boldsymbol{min}})$

e) Independence of Linear Transformation: For any linear scale transformation Ψ, $\Psi(\Phi(\mathcal{S}, \boldsymbol{R}^{\boldsymbol{min}})) = \Phi(\Psi(\mathcal{S}, \boldsymbol{R}^{\boldsymbol{min}}))$.

f) Symmetry: If $\mathcal{S}$ is invariant under all exchanges of users, $\Phi_j(\mathcal{S}, \boldsymbol{R}^{\boldsymbol{min}}) = \Phi_{j'}(\mathcal{S}, \boldsymbol{R}^{\boldsymbol{min}}), \forall j, j' \in \mathcal{N}$.

Axioms a), b) and c) give the definition of bargaining set, and axioms d), e) and f) are called axioms of fairness. Further, Theorem 1 shows the existence and uniqueness of NBS that satisfies the above axioms.

Theorem 1: There exists an unique solution $\Phi(\mathcal{S}, \boldsymbol{R}^{\boldsymbol{min}})$ *that satisfies all axioms in definition 2, and the solution satisfies*

$$\Phi\left(\mathcal{S}, \boldsymbol{R}^{\boldsymbol{min}}\right) \in \arg_{\bar{\boldsymbol{R}} \in \boldsymbol{S}} \max \prod_{i=1}^{N} \left(\bar{R}_i - R_i^{min}\right). \tag{11}$$

Proof: The similar detailed proof can be found in [15]. □

Therefore, we formulate the resource allocation optimization problem as

$$\max_{a_{i,k}, p_{i,k}} \prod_{i=1}^{N} \left(\sum_{k=1}^{K} a_{i,k} w_0 \log_2 \left(1 + \frac{p_{i,k} g_{i,k}}{\sigma_k^2}\right) - R_i^{\min} \right) \tag{12}$$

$$\text{s.t. } R_i \geq R_i^{\min} \tag{13}$$

$$\sum_{k=1}^{K} a_{i,k} p_{i,k} \leq P_i^{\max} \tag{14}$$

$$\sum_{k=1}^{K} a_{i,k} \leq 1 \qquad \forall k \tag{15}$$

$$a_{i,k} \in \{0, 1\} \qquad \forall i, k. \tag{16}$$

where the optimization objective reflects the generalized proportional fairness because it is beneficial to the user with less tolerated rate $R_i^{\min}$ and one user's performance is unchanged from the other user's channel conditions [3]. Constraint (13) indicates that the data transmission rate of user i is larger than the minimized rate R_i^{min} that user i can tolerate. Constraint (14) is the transmission power constraint, i.e., the total power spent on all subchannels occupied for input data transmission is no more than the maximum power that user i can supply. Constraint (15) states that each subchannel can only be occupied by one user at a time.

III. SOLUTION OF NASH BARGAINING GAME BASED RESOURCE ALLOCATION

The NBS optimization problem in (12)-(16) is a mixed integer nonlinear programming problem (MINLP), which is NP-hard. The optimal solution can be obtained by exhaustive search. However, with the increasing number of users and subchannels, the MINLP optimization problem has a high complexity. Therefore, in order to reduce the computation complexity, we adopt the time-sharing relaxation [16] to transform the MINLP problem into a nonlinear real-number programming problem. The time-sharing method has been widely used to resolve nonlinear combinational optimization problems for multi-user subchannel allocation in OFDMA systems [17].

We introduce allocation time variable τ_i, which means the fraction of time when user i occupies the all subchannels of the BS. We assume that $T^{\max} = \max\{T_i^{\max}\}$ is the maximum time delay that satisfies all users' time constraints, then the optimal amount of time allocated to user i is $\tau_i^* T^{\max}$. Now, we can transform the resource allocation optimization problem into the following optimization problem

$$\max_{\tau_i, p_{i,k}} \prod_{i=1}^{N} \left(\tau_i \sum_{k=1}^{K} w_0 \log_2 \left(1 + \frac{p_{i,k} g_{i,k}}{\sigma_k^2}\right) - R_i^{\min} \right) \tag{17}$$

$$\text{s.t. } \tau_i \sum_{k=1}^{K} w_0 \log_2 \left(1 + \frac{p_{i,k} g_{i,k}}{\sigma_k^2}\right) \geq R_i^{\min} \tag{18}$$

$$\sum_{k=1}^{K} p_{i,k} \leq P_i^{\max} \tag{19}$$

$$\sum_{i=1}^{N} \tau_i \leq 1. \tag{20}$$

Condition (20) is the normalized time interval constraint for all the users. Then the previous discrete problem (12)-(16) is transformed into a continuous problem (17)-(20) in which the continuous variables τ_i and $p_{i,k}$ are optimized.

A. SOLUTION OF OPTIMAL POWER ALLOCATION

From [11], we know that necessary and sufficient condition for the optimal allocation solution $p_{i,k}^*$ exists, if and only if $p_{i,k}^*$ is the optimal solution of the following optimization problem

$$\max_{p_{i,k}} w_0 \log_2\left(1+\frac{p_{i,k}g_{i,k}}{\sigma_k^2}\right). \tag{21}$$

$$\text{s.t.} \sum_{m=1}^{K} p_{i,k} \le P_i^{\max} \tag{22}$$

Obviously, optimization function of (21) is log-concave with respect to $p_{i,k}$, and constraint (22) is linear. Thus the optimization problem in (21) and (22) is convex. The Lagrangian function is given as follows

$$L_1(p_{i,k},\lambda) = \sum_{k=1}^{K} w_0 \log_2\left(1+\frac{p_{i,k}g_{i,k}}{\sigma_k^2}\right) - \lambda\left(\sum_{k=1}^{K} p_{i,k} - P_i^{\max}\right), \tag{23}$$

where λ is the Lagrangian multiplier.

Based on the KKT condition [18], the following conditions must be satisfied

$$\begin{cases} \dfrac{\partial L}{\partial p_{i,k}} = 0 \\ \lambda\left(\displaystyle\sum_{k=1}^{K} p_{i,k} - P_i^{\max}\right) = 0 \\ \lambda \ge 0. \end{cases} \tag{24}$$

Then, we can obtain the optimal power allocation $p_{i,k}^*$ for user i on subchannel k, which is given as

$$p_{i,k}^* = \left(\frac{1}{K}\left(P_i^{\max} + \sum_{k=1}^{K}\frac{\sigma_k^2}{g_{i,k}}\right) - \frac{\sigma_k^2}{g_{i,k}}\right)^+ \tag{25}$$

where $x^+ \equiv \max(0,x)$.

B. SOLUTION OF OPTIMUM TIME ALLOCATION

From (17), we know that it is necessary that $p_{i,k}^*$ must satisfy the following inequality (26)

$$\sum_{k=1}^{K} w_0 \log_2\left(1+\frac{p_{i,k}^* g_{i,k}}{\sigma_k^2}\right) \ge R_i^{\min} \tag{26}$$

then the transformed continuous optimization problem (17)-(20) has a solution.

Since the concave and monotonic property of logarithm function, we take the logarithm of (26). With the given $p_{i,k}^*$, we can get the following optimization problem

$$\max_{\tau_i} \sum_{i=1}^{N} \ln\left(\tau_i \sum_{k=1}^{K} w_0 \log_2\left(1+\frac{p_{i,k}^* g_{i,k}}{\sigma_k^2}\right) - R_i^{\min}\right) \tag{27}$$

$$\text{s.t.} \sum_{i=1}^{N} \tau_i \le 1. \tag{28}$$

The Lagrange function is formulated as

$$Q(\tau_i,\theta) = \sum_{i=1}^{N} \ln\left(\tau_i \sum_{k=1}^{K} w_0 \log_2\left(1+\frac{p_{i,k}g_{i,k}}{\sigma_k^2}\right) - R_i^{\min}\right) - \theta\left(\sum_{i=1}^{N}\tau_i - 1\right), \tag{29}$$

where θ is the Lagrange multiplier. Employing KKT conditions, then we have

$$\begin{cases} \dfrac{\partial L}{\partial \tau_i} = 0 \\ \theta\left(\displaystyle\sum_{i=1}^{N} \tau_i - 1\right) = 0 \\ \theta \ge 0. \end{cases} \tag{30}$$

The optimization time allocation τ_i^* for user i is calculated as

$$\tau_i^* = \frac{1}{N}\left(1 - \sum_{i=1}^{N}\frac{R_i^{\min}}{R_i^*}\right) + \frac{R_i^{\min}}{R_i^*}. \tag{31}$$

C. SUBCHANNEL MATCHING AND USER PRIORITY TRANSMISSION ALGORITHM

After obtaining the optimal time fraction allocation, we develop an algorithm based on Hungarian method and two-band partition strategy [3]. This algorithm can ensure mobile users meeting their transmission delay constraints by adjusting the user priority. The algorithm is presented in Algorithm 1.

First, we determine user priority by sorting all the mobile users in ascending order of time delay constraint, i.e., $\forall i,j \in \mathcal{N}$ and $T_j^{\max} \le T_i^{\max}$, we have $j < i$. For the case $T_j^{\max} = T_i^{\max}$, if $\tau_j < \tau_i$, then $j < i$; Then, after sorting, the user with less ordinal in $\mathcal{N}$ has the higher priority for input data transmission.

Given the order for transmission, we will check whether each user i can meet its time delay constraint requirement $T_i^{\max}$ or not. For any user i, let

$$\delta_i = T_i^{\max} - \left(\sum_{j=1}^{i-1}\tau_j T^{\max} + \tau_i T^{\max}\right) \tag{32}$$

where $\sum_{j=1}^{i-1}\tau_j T^{\max}$ is the time cost by mobile users with higher priority than user i for transmission. If $\delta_i > 0$, which means user i not only meets its time delay constraint but has time left after transmission within time duration $\tau_i T^{\max}$. And for these users, we group them into set $\mathcal{Q}$. Otherwise, for any user i with $\delta_i < 0$, which means it cannot meet its time delay constraint, then it belongs to set $\mathcal{M}$. And each user $m \in \mathcal{M}$ calculates its data size η_m that can not be transmitted within its time constraint $T_m^{\max}$, which is expressed as $\eta_m = d_m - \tau_m T_m^{\max} R_m$.

Algorithm 1 Subchannel Matching and User Priority Transmission Algorithm

```
Input: mobile user delay constrain $T_i^{\max}$, time fraction $\tau_i$, maximum time delay $T^{\max}$, data size $d_i$.
order users such that $T_j^{\max} \le T_i^{\max}$, $\forall i, j \in \mathcal{N}$ and $j < i$;
let $\delta_i = T_i^{\max}$ - ($\sum_j^{i-1} \tau_j T^{\max} + \tau_i T^{\max}$), $\mathcal{Q} \leftarrow \emptyset$, $\mathcal{M} \leftarrow \emptyset$;
for $i = 1, \ldots, N$ do
    if $\delta_i > 0$ then $\mathcal{Q} \leftarrow \mathcal{Q} \cup i$;
    else if $\delta_i < 0$ then $\mathcal{M} \leftarrow \mathcal{M} \cup i$;
    end if
end for
if $\mathcal{M} = \emptyset$ then
    goto loop;
end if
for $m = \{1, \ldots, M\} \in \mathcal{M}$ do
    let $\eta_m = d_m - \tau_m T_m^{\max} R_m$;
end for
user $i \in \mathcal{N}$ constructs its subchannel list on SNR for all $k \in \mathcal{K}$ in descending order, $\forall i$;
if user $q \in \mathcal{Q}$ and user $m \in \mathcal{M}$ successfully matched by the Hungarian algorithm then
    let $x_{q,m} = 1$;
else
    let $x_{q,m} = 0$;
end if
loop:
for $i = \{1, \ldots, N\ \} \in \mathcal{N}$ do
    if user $i \in \mathcal{Q}$ then
        user $i$ shares subchannels with user $m \in \mathcal{M}$ by two-band partition algorithm, if $x_{i,m} = 1$;
        user $i$ occupies the first $\kappa_i$ subchannels on its channel list satisfying $R_i^{\kappa_i} \ge d_i/(\tau_i T^{\max}) > R_i^{\kappa_i - 1}$;
        number of subchannels user $m$ occupied is $K - \kappa_i$;
        user $i$ and $m$ start transmission on its own subchannels until user $i$ finish transmission;
        let $\eta_m \leftarrow \max(\eta_m - \tau_i T^{\max} R_m^{K-\kappa_i}, 0)$;
        if $\eta_m = 0$ then $M \leftarrow M - m$;
        end if
    else
        user $i$ starts to transmit with time limit $\tau_i T^{\max}$;
    end if
end for
```

Then, we discuss the following three different cases: 1) $\mathcal{M} \subseteq \emptyset$; 2) $\mathcal{Q} \subseteq \emptyset$, $\mathcal{M} \nsubseteq \emptyset$ and 3)$\mathcal{Q} \nsubseteq \emptyset$, $\mathcal{M} \nsubseteq \emptyset$. In the first case, all users in $\mathcal{N}$ meet their time delay constraint, then they transmit their input data in order. For the case $\mathcal{Q} \subseteq \emptyset$ and $\mathcal{M} \nsubseteq \emptyset$, it won't occur. And we give a proof for that by contradiction. We know that all the input data transmission of users in $\mathcal{N}$ can be accomplished within time $T^{\max}$. As $\mathcal{M} \nsubseteq \emptyset$, then there must be user i which has time left after transmission within its time duration $\tau_i T^{\max}$ such that $\mathcal{Q} \nsubseteq \emptyset$. Next, we will design a negotiation strategy to improve bandwidth efficiency and meet all users' time delay constraint requirement in the third case.

Then, for any user $q \in \mathcal{Q}$, we choose a user $m \in \mathcal{M}$ to match with user q with the aim of maximizing $(R_q - R_q^{\min})(R_m - R_m^{\min})$ by Hungarian algorithm. If the match succeeds, we let $x_{q,m} = 1$. Otherwise, $x_{q,m} = 0$. After the match, the user i in $\mathcal{N}$ transmits its data in the order according to its priority. If user $i \in \mathcal{Q}$ and $x_{q,m} = 1$, when it's the turn for user i to transmit data, user i shares subchannels with user m by two-band partition algorithm. In this case, user i occupies the first κ_i subchannels on its channel list, and user $m \in \mathcal{M}$ occupies the rest $K - \kappa_i$ subchannels. Then both user i and q begin to transmit their data simultaneously on its subchannels, respectively, until user i finish transmission. In the iteration, if user m does not accomplish the transmission of data size η_m, it will wait the next iteration to continue the rest transmission when it can share the subchannels with other users in $\mathcal{Q}$.

The total complexity of the Hungarian algorithm is $O(N^2)$, and the complexity of the two-band partition algorithm for all users in $\mathcal{N}$ is $O(N^2 log_2 N)$. Therefore, from Algorithm 1, we can obtain the complexity of the proposed algorithm is $O(N^2 log_2 N)$.

IV. PERFORMANCE EVALUATION

In this section, we evaluate the performance of proposed method in terms of system throughput and fairness. We introduce the simulation setup at first. Then extensive simulations are provided and analyzed. A brief discussion is provided to highlight the performance of the proposed method.

A. SIMULATION SETUP

We consider a group of mobile users that are randomly deployed in the OFDMA wireless network with MEC. They are served by one base station with the coverage radius of 200 m. Two scenarios are considered in the simulations. In both scenarios, the system parameters are set as follows unless otherwise specified. The number of subchannels $K = 180$, and the bandwidth of each subchannel $w_0 = 20$ kHz. The maximum transmission power of mobile users is set to 20 mW. The noise power of each subchannel k, σ_k^2, is 10^{-11} mW. The two scenarios are described as follows:

Scenario 1: Three users are considered in Scenario 1, labeled as users 1, 2 and 3. The three users have the same minimum rate requirement, i.e., $R_1^{\min} = R_2^{\min} = R_3^{\min} = 0.7$ Mbps. The distances from user 1 and user 2 to the base station are fixed at $D_1 = 100$ m and $D_2 = 80$ m, respectively, while the distance between user 3 and the base station, D_3, varies from 100 m to 200 m. Then, we can evaluate if the time fraction of subchannels can be dynamically allocated when D_3 increases. The propagation loss factor is set to 3. The simulation results of this scenario are provided in Fig. 2 and Fig. 3.

Scenario 2: In order to evaluate the superiority of the proposed method over existing methods in maximizing the

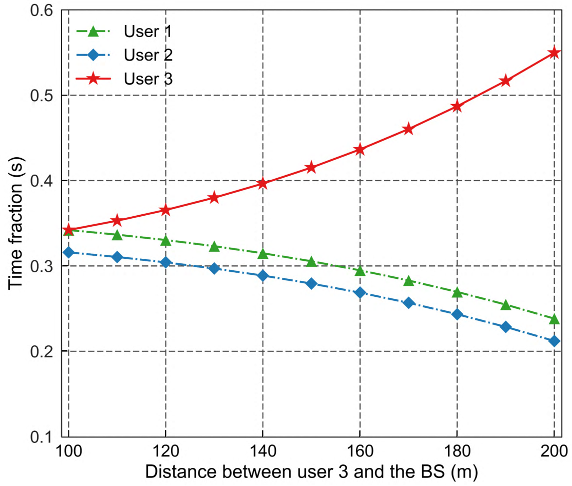


FIGURE 2. **The time fraction allocation of users when user$\dot{3}$ moves away from the base station with D_3 increasing.**

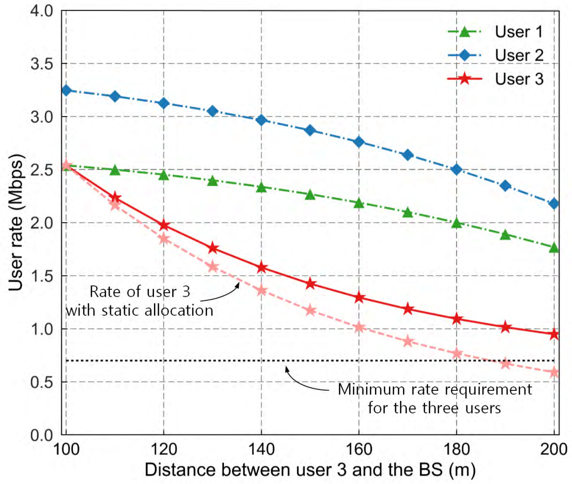


FIGURE 3. **The channel rate of users when user$\dot{3}$ moves away from the base station with D_3 increasing.**

overall system rate and guaranteeing the fairness, we will consider a more general case where the number of users increases from 10 to 50. The users are randomly located within the coverage radius of the base station, then the distance D_i between user i and the base station is no larger than 200 m. The minimum rate requirement for user i is $R_i^{\min} \in [0.3, 0.8]$. Then the proposed Nash Bargaining method is compared with the classical equal allocation method and max-min method [19] in terms of system overall rate and fairness. The simulation results of this scenario are provided in Fig. 4 and Fig. 5.

B. SIMULATION RESULTS

The proposed method is evaluated through three performances: effectiveness, fairness and throughput. The effectiveness means that any user can meet the minimum rate requirement by negotiating with other users. The fairness index is the proportional fairness considering the individual

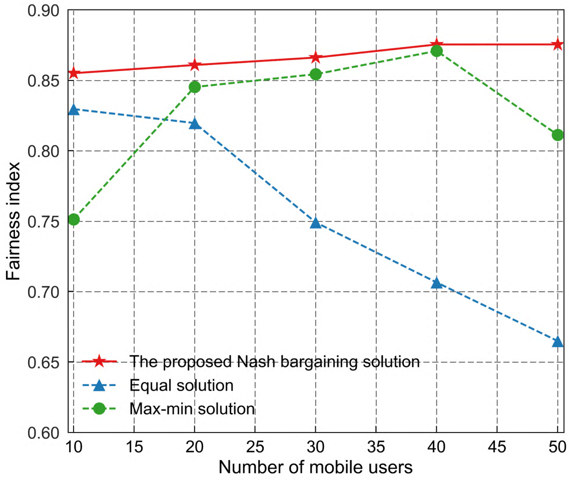


FIGURE 4. **The comparison between the proposed method with existing methods on fairness.**

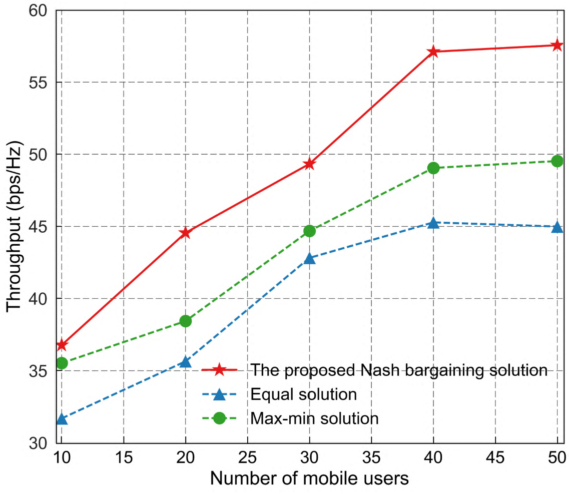


FIGURE 5. **The comparison between the proposed method and existing methods on throughput.**

minimum rate. The throughput is the system total rate reflecting the resource utilization. The effectiveness evaluation is conducted on scenario 1. The fairness and throughput evaluations are conducted on scenario 2.

1) EFFECTIVENESS

We evaluate the effectiveness of the proposed method in dynamically allocating the time fraction of subchannels with the setting of Scenario 1. In this scenario, user 3 moves away from the base station, resulting in the transmission rate degradation of user 3. The effectiveness of proposed method is evaluated if the time fraction of subchannels can be dynamically allocated.

Fig. 2 shows the time fraction allocation of subchannels for the three users, where we find the time fraction of user 3 increases as user 3 moves away from the base station. This is because that when the communication distance of user 3 increases, the channel transmission rate degrades due to the path loss effect. In order to guarantee the minimum rate requirement of user 3, the time fraction of user 3 increases from 0.34 to 0.55 to mitigate the channel rate degradation caused by the increase of communication distance. From Fig. 2, it can be seen that the rates of user 1 and user 2 both decrease because some of their time fractions are allocated to user 3 so that it can meet the minimum rate requirement. The rate reduction can be accepted by these users because their rates are still greater than their minimum rates.

Fig. 3 shows the evolution of three users' rates when user 3 moves away from the base station. It is shown that the rate of user 3 decreases during the movement due to the channel gain degradation. But the channel rate is still larger than the minimum rate requirement (0.7 Mbps). This is because the increasing time friction, just as shown in Fig. 2, compensates the channel rate degradation, which satisfies the minimum rate requirement. For the static allocation, however, the channel rate of user 3 falls below the threshold since it cannot react to the change of channel conditions. From Fig. 3, we find that the rates of user 1 and user 2 decrease because the time fractions of the two users decrease as shown in Fig. 2. The rates of the users still satisfy the minimum rate requirement as minimum rate is explicitly considered in the optimization problem.

2) FAIRNESS

We evaluate the performance of the proposed method in guaranteeing the fairness of users with the setting of Scenario 2. We first introduce the mathematical expression of the fairness index, then we compare the fairness index of the proposed method with existing methods.

The fairness index γ can be mathematically defined as [20]

$$\gamma = \frac{\left(\sum_{i=1}^{N} \left(R_i/R_i^{\min}\right)\right)^2}{\left(N \cdot \left(\sum_{i=1}^{N} \left(R_i/R_i^{\min}\right)^2\right)\right)}, \tag{33}$$

where $\gamma \in (0, 1]$ characterizes if all users can satisfy the minimum rate requirement. If $\gamma = 1$, the ratios of users' rates to the corresponding minimum rate are the same, which implies that the resource allocation is perfectly fair.

Fig. 4 shows the comparison of the proposed method with existing methods in fairness. From Fig. 4, we find that the fairness of the proposed method outperforms the equal allocation method and max-min allocation method, especially when the number of users increases. This is because the time fraction can be allocated dynamically to meet each user's requirement. In the equal allocation method, the time fraction is allocated equally, which implies that the increase of user number results in less resources for each user. This leads to a larger rate deviation among users due to the different channel conditions, and then the fairness index decreases with the number of users. The max-min allocation method achieves a certain degree of fairness because it first satisfies the requirement of the worst user, which means that the disparity is restrained. But when the number of users is large enough, the users with higher resource can not be satisfied gradually, implying that the fairness index begins to drop. The proposed method provides a better fairness index due to its proportional allocation based on the minimum rate requirement of each user.

3) THROUGHPUT

We evaluate the performance of the proposed method in maximizing the system throughput with the setting of Scenario 2. The throughput of the proposed method is compared with those of the equal allocation method and max-min allocation method. In Fig. 5, it is shown that the throughput of the proposed method is much larger than those of the other methods. This is because that in the proposed method, resources are allocated proportionally based on the user's requirement and the channel condition simultaneously, which takes the full advantage of channels with high SNR. For the max-min allocation method, the rate of the worst user is maximized, while the other ones with better channel gains are penalized, which degrades the system throughput. In the equal allocation method, the system throughput is also degraded since neither the channel condition nor minimum rate requirement is considered in the design.

V. CONCLUSION

In this paper, we investigate the overall network throughput maximization problem in MEC to meet the resource requirements of ever-increasing mobile devices. As a cooperative game theory, the Nash bargaining game is adopted to ensure the fairness of resource allocation. Furthermore, the existence and uniqueness of the Nash bargaining game based solution has been investigated. And an algorithm is developed to determine user priority by considering the users' delay constraint. Evaluation results demonstrate that our scheme can improve the overall system rates considerably, and ensure the fairness among various users. In the future work, the coalition of MECs will be studied to improve the resources utilization further.

REFERENCES

[1] Y. Mao, C. You, J. Zhang, K. Huang, and K. B. Letaief, "A survey on mobile edge computing: The communication perspective," *IEEE Commun. Surveys Tuts.*, vol. 19, no. 4, pp. 2322–2358, 4th Quart., 2017.

[2] 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.

[3] Z. Han, Z. Ji, and K. J. R. Liu, "Fair multiuser channel allocation for OFDMA networks using nash bargaining solutions and coalitions," *IEEE Trans. Commun.*, vol. 53, no. 8, pp. 1366–1376, Aug. 2005.

[4] Y. Zhao, X. Fang, R. Huang, and Y. Fang, "Joint interference coordination and load balancing for OFDMA multihop cellular networks," *IEEE Trans. Mobile Comput.*, vol. 13, no. 1, pp. 89–101, Jan. 2014.

[5] C. Singhal, S. Kumar, S. De, N. Panwar, R. Tonde, and P. De, "Class-based shared resource allocation for cell-edge users in OFDMA networks," *IEEE Trans. Mobile Comput.*, vol. 13, no. 1, pp. 48–60, Jan. 2014.

[6] P. Zhao, H. Tian, C. Qin, and G. Nie, "Energy-saving offloading by jointly allocating radio and computational resources for mobile edge computing," *IEEE Access*, vol. 5, pp. 11255–11268, Jun. 2017.

[7] J. Ren, G. Yu, Y. Cai, and Y. He, "Latency optimization for resource allocation in mobile-edge computation offloading," *CoRR*, Apr. 2017. [Online]. Available: http://arxiv.org/abs/1704.00163

[8] X. Chen, L. Jiao, W. Li, and X. Fu, "Efficient multi-user computation offloading for mobile-edge cloud computing," *IEEE/ACM Trans. Netw.*, vol. 24, no. 5, pp. 2795–2808, Oct. 2016.

[9] X. Ma, C. Lin, X. Xiang, and C. Chen, "Game-theoretic analysis of computation offloading for cloudlet-based mobile cloud computing," in *Proc. 18th ACM Int. Conf. Modeling, Anal. Simulation Wireless Mobile Syst. (MSWiM)*, New York, NY, USA, 2015, pp. 271–278. [Online]. Available: http://doi.acm.org/10.1145/2811587.2811598

[10] X. Xu, J. Liu, and X. Tao, "Mobile edge computing enhanced adaptive bitrate video delivery with joint cache and radio resource allocation," *IEEE Access*, vol. 5, pp. 16406–16415, Aug. 2017.

[11] K. D. Lee and V. C. M. Leung, "Fair allocation of subcarrier and power in an OFDMA wireless mesh network," *IEEE J. Sel. Areas Commun.*, vol. 24, no. 11, pp. 2051–2060, Nov. 2006.

[12] Y. Mao, J. Zhang, and K. B. Letaief, "Dynamic computation offloading for mobile-edge computing with energy harvesting devices," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 12, pp. 3590–3605, Dec. 2016.

[13] H. Yaiche, R. R. Mazumdar, and C. Rosenberg, "A game theoretic framework for bandwidth allocation and pricing in broadband networks," *IEEE/ACM Trans. Netw.*, vol. 8, no. 5, pp. 667–678, Oct. 2000.

[14] Q. Ni and C. C. Zarakovitis, "Nash bargaining game theoretic scheduling for joint channel and power allocation in cognitive radio systems," *IEEE J. Sel. Areas Commun.*, vol. 30, no. 1, pp. 70–81, Jan. 2012.

[15] D. Fudenberg and J. Tirole, *Game Theory*, vol. 1. Cambridge, MA, USA: MIT Press, pp. 841–846, 1993.

[16] C. Y. Wong, R. S. Cheng, K. B. Lataief, and R. D. Murch, "Multiuser OFDM with adaptive subcarrier, bit, and power allocation," *IEEE J. Sel. Areas Commun.*, vol. 17, no. 10, pp. 1747–1758, Oct. 1999.

[17] M. Tao, Y.-C. Liang, and F. Zhang, "Resource allocation for delay differentiated traffic in multiuser OFDM systems," *IEEE Trans. Wireless Commun.*, vol. 7, no. 6, pp. 2190–2201, Jun. 2008.

[18] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge, U.K.: Cambridge Univ. Press, 2004.

[19] W. Rhee and J. M. Cioffi, "Increase in capacity of multiuser OFDM system using dynamic subchannel allocation," in *Proc. IEEE 51st Veh. Technol. Conf. (VTC-Spring)*, vol. 2. May 2000, pp. 1085–1089.

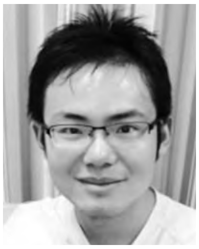
[20] R. Jain, D. Chiu, and W. Hawe, "A quantitative measure of fairness and discrimination for resource allocation in shared computer systems," DEC Res. Rep. TR-301, Sep. 1984.

ZHENGFA ZHU received the bachelor's degree in communication engineering from Central South University, Changsha, China, in 2009, where he is currently pursuing the Ph.D. degree with the School of Information Science and Engineering. His current research interests include cloud computing, mobile edge computing, and wireless networks.

JUN PENG (M'08) received the B.S. degree from Xiangtan University, Xiangtan, China, in 1987, and the M.Sc. degree from the National University of Defense Technology, Changsha, China, in 1990, and the Ph.D. degree from Central South University, Changsha, in 2005. In 1990, she joined the staff of Central South University. From 2006 to 2007, she was a Visiting Scholar with the School of Electrical and Computer Science, University of Central Florida, Orlando, FL, USA. She is currently a Professor with the School of Information Science and Engineering, Central South University. Her research interests include cooperative control, cloud computing, and wireless communications.

XIN GU received the bachelor's degree in communication engineering from Central South University, Changsha, China, in 2015, where she is currently pursuing the Ph.D. degree with the School of Information Science and Engineering. Her current research interests include cloud computing, mobile edge computing, and wireless networks.

HENG LI (M'17) received the bachelor's and Ph.D. degrees from Central South University, Changsha, China, in 2011 and 2017, respectively. From 2015 to 2017, he was a Research Assistant with the Department of Computer Science, University of Victoria, Victoria, BC, Canada. He is currently an Assistant Professor with the School of Information Science and Engineering, Central South University. His current research interests include cooperative control and cyber physical systems.



KAIYANG LIU received the B.S. degree from the School of Information Science and Engineering, Central South University, Changsha, China, in 2012, where he is currently pursuing the Ph.D. degree. His general research interests include the broad area of wireless communication and computer networking, with special emphasis on resource management and scheduling in cloud computing systems, mobile sensor networks, and network optimization.

ZHUOFU ZHOU received the bachelor's and master's degrees in communication engineering from Central South University, Changsha, China, in 2011 and 2014, respectively. His current research interests include mobile cloud computing and big data.

WEIRONG LIU (M'14) received the B.S. and M.S. degrees from the School of Information Science and Engineering, Central South University, Changsha, China, and the Ph.D. degree from the Laboratory of Complex Systems and Intelligence Science, Institute of Automation, Chinese Academy of Sciences, Beijing, China. He is currently an Associate Professor with the School of Information Science and Engineering, Central South University. His special fields of interests include cooperative control, nonlinear control, wireless sensor network, and embedded systems.