Multi-MEC server multi-user resource allocation in heterogeneous network

ZhenQin³, Yuanxiu Liao¹,²,³,* and Dejin Shen³

¹Guangxi Key Lab of Multi-Source Information Mining and Security,
²Guangxi Regional Multi-Source Information Integration and Intelligent Processing Collaborative Innovation Center
³Faculty of Computer Science and Information Engineering, Guangxi Normal University, Guilin, Guangxi 541004, China.

*Corresponding author’s e-mail: liaoyuanxiu@mailbox.gxnu.edu.cn

Abstract. Aiming at the coexistence of large and small cells in heterogeneous cellular networks, the difference of computing resources between mobile equipment and base stations, and the diversity of service quality demands of users, a resource allocation algorithm based on matching game theory was proposed. We assume that some users have computing needs and can perform MEC offloading or D2D offloading. We define cost performance as the ratio of the data rate per user to its price cost, and resource allocation as the cost performance maximization problem. Considering that the formulation problem is a mixed integer nonlinear problem and it is difficult to obtain the optimal solution, we design an improved Gale-Shapley algorithm based on matching game theory to obtain the feasible solution of the problem. Simulation results show that the proposed multi-MEC server multi-user resource allocation strategy can effectively reduce user service delay and improve system performance in heterogeneous cellular networks.

1. Introduction

Due to the limited computing, storage capacity and battery power of mobile equipment, they cannot support applications with high computing capacity and low latency demands. Therefore, complex computing tasks need to be offloaded from base stations to cloud computing center for execution. However, the migration of computing tasks to cloud computing centers for execution will bring about a large amount of data transmission. Moreover, limited wireless spectrum resources, low transmission power of mobile devices and long-distance data transmission not only lead to the increase of transmission delay, but also increase the load of cloud computing center, which will bring serious impact on delay sensitive users. To this end, 5G combines mobile edge computing (MEC) technology with MEC servers deployed on each small base station to alleviate network load, reduce transmission delay and improve user satisfaction.

Although the combination of 5G and mobile edge computing technology brings many advantages, the resources at the network edge are limited. Therefore, resource allocation needs to be reasonably optimized according to user needs to maximize system performance while ensuring user service quality.

In recent years, the problem of resource allocation has attracted the attention of many researchers. Many researchers have focused on the allocation of wireless and computing resources in a
heterogeneous network with multiple users and multiple MEC servers [1-3]. In [1], the author aims to minimize the computational cost of the system in terms of computing time and energy consumption, and decomposes the joint resource allocation problem into the user offloading transmission power sub-problem and the MEC server computing resource allocation sub-problem. First, the inter-cell interference is approximated by using the dichotomy to find the transmission power of the offloaded user, and then the KKT optimality condition is used to solve the calculation resource allocation problem. In [2], the author additionally considers user costs. And under the condition that the user delay constraint is satisfied, the communication cost of the mobile terminal is minimized by optimizing the uplink transmission power and computing resource scheduling. In [3], the author studied the joint sub-channel and power allocation optimization sub-problem under the constraints of inter-user interference and energy consumption, with the goal of minimizing the weighted sum of user delays. In addition, a low-complexity algorithm using CVX tools and continuous convex approximation methods is proposed, and interference management is considered, and cross-layer interference is managed by setting interference thresholds.

In recent years, matching game theory has been widely used to study resource allocation in heterogeneous cellular networks. The matching game framework can provide a better model to characterize the interaction between different participants, can appropriately present the preferences of the system requirements, and provide feasible solutions. Many researchers have studied the problem of resource allocation with the goal of minimizing the total energy consumption of users [4-5]. In [4], the author uses NOMA technology to improve the throughput of cellular networks, and proposes a joint optimization problem involving user association, computing resource allocation, and transmission power control. In [5], the author jointly optimized the allocation of wireless and computing resources. The computing resource and power control are modeled as a many-to-one matching, and the solution is solved by establishing a list of preferences between users and resources. In [6], the author aims to maximize the transmission rate, uses matching game theory to solve the problem, and improves the Gale-Shapley algorithm by adding interference management. In [7], the author studied the problem of computing resource allocation in a three-layer fog computing network, with the goal of maximizing cost-effectiveness. The pairing between the three types of participants is modeled as a pairwise many-to-one matching, and the stable matching between the three is solved through iteration. But the problem of wireless resource allocation is not studied. The author studied the joint wireless and computing resource allocation problem under the three-layer fog computing network in [8] to optimize system performance and user satisfaction. The user-fog node-service provider is modeled as an SPA resource allocation model, and the matching game SPA(s, p) algorithm is used to find a stable match between users and resources. At the same time, the UOC (user-oriented cooperation) strategy is proposed to remove the instability caused by external effects to ensure the stability of the system. Although the above studies have used matching game theory to solve resource allocation problems, they also considered the interests of users and service providers. But it ignores the computing power of idle devices in the network. Using mobile devices in an free state in the network as a small server to provide services for users with computing needs can make full use of the computing power in the network.

In this article, we study resource allocation strategies based on matching games and propose a three-layer resource allocation model under 5G heterogeneous cellular networks. In order to make full use of the computing power of idle devices, and to encourage the enthusiasm of idle devices to provide services through a compensation mechanism. The computing resources required for user tasks can be provided by the MEC server deployed on the small cell, or can be provided by free equipment. Then define the multi-user resource allocation optimization problem with the goal of maximizing cost performance, and finally design a resource allocation algorithm based on matching game.

2. System model

As shown in Figure 1, We consider a heterogeneous cellular network consisting of one base station (MBS), multiple small base station (SBS) and mobile equipment. A MEC server is deployed on each
BS, the MBS has K orthogonal channels, which is expressed as $K = \{c_1, c_2, ..., c_K\}$, and L FBS under the coverage of the macro station, which is expressed as $S = \{s_1, s_2, ..., s_L\}$, SBS and MBS share K channels. Mobile equipment with a computing need is called user, denoted as $U = \{u_1, u_2, ..., u_N\}$. The other part of the mobile equipment that is free and can provide computing services to users is called free equipment and is represented as $RU = \{ru_1, ru_2, ..., ru_M\}$. Each user has computation-intensive and delay-sensitive tasks, and the computing resources needed by the user are provided by the SBS or free equipment, that is, both the SBS and free equipment can act as service providers. In order to improve channel utilization, it is assumed that each channel can be shared by multiple users, and there will be interference between users sharing the channel.

For delay-sensitive users, they are willing to pay the service provider based on the size of the requested task and the latency requirements. We can see the resource allocation between the user and the service provider as a mapping between the user and the service provider, represented in binary, where $\mu_{u_i,r u_m} = 1$ if The user $u_i$ offloads the task to the free equipment $ru_m$ via channel $c_k$ and $\mu_{u_i,r u_m} = 0$ otherwise, where $\gamma_{u_i,s_j} = 1$ if The user $u_i$ offloads the task to the base station $s_j$ via channel $c_k$, and $\gamma_{u_i,s_j} = 0$. Next, discuss the communication model in the network. It is assumed that user $u_i$ carries a data size of $D_i$ (bit), which corresponds to the CPU cycle rate required for the processing task. For simplicity of calculation, it is assumed that the relationship between $D_i$ and $DC_i$ is linear [9].

![Figure 1. system model](image)

2.1. Communication between the user and the base station

When user $u_i$ offloads the task to base station $s_j$ via channel $c_k$. We define the SNR received from the user $u_i$ to the base station $s_j$ using channel $c_k$ as follows.

$$\Gamma_{u_i,s_j}^{c_k} = \frac{P_{u_i}g_{u_i,s_j}^{c_k}}{\sum_{u_i \in U, ru_m \in RU} \mu_{u_i,ru_m}^{c_k}P_{u_i}g_{u_i,ru_m} + \sum_{u_i \in U, s_j \notin S} \gamma_{u_i,s_j}^{c_k}P_{u_i}g_{u_i,s_j}^{c_k} + \sigma^2}$$

where $P_{u_i}$ and $g_{u_i,s_j}^{c_k}$ are the transmission power and channel gain between user $u_i$ and base station $s_j$ using channel $c_k$, $g_{u_i,ru_m}^{c_k}$ represents channel interference gain from any other user $u_i$ and base station $s_j$ using channel $c_k$, $g_{u_i,ru_m}$ represents channel interference gain from any other user $u_i$ and free...
equipment \( r_u \) using channel \( c_k \). In order to ensure the quality of data transmission, we ask SNR \( \Gamma_{u_i, s_j}^{c_k} \) is higher than threshold \( \Gamma_{\text{min}} \).

The data transmission rate from user \( u_i \) using channel \( c_k \) to offload the task to \( s_j \), can be expressed as follows.

\[
R_{u_i, s_j}^{c_k} = B_c \log_2 (1 + \Gamma_{u_i, s_j}^{c_k})
\] (2)

Assuming that the resources of a base station can be shared by multiple users, the capacity of a base station can accommodate at most \( Q^{s_j} \) users, the computing capacity of SBS \( s_j \) is \( C_j \), and the computing capacity of base station \( s_j \) assigned to user \( u_i \) is \( c_i^j \). Assuming that the small base station allocates equal computing power to each user, noted as \( c_i^j = \frac{c_i}{\sum_{u\in U} \mu_{u_i, s_j}} \). An important factor that users care about when choosing resources is service latency, and users expect shorter service latency. Total service delay \( T_{u_i, s_j}^k \) consists of transmission delay \( t_t \), processing delay \( t_p \), and return delay \( t_r \). The return link transmission delay \( t_r \) is often ignored when calculating the total service delay because the output data size of a computed task is much smaller than the input data, so the delay in transferring the output data from the free equipment to the user is much smaller than the transmission delay of the input data and is therefore ignored[10]. So we define the service delay as follows.

\[
T_{u_i, s_j}^k = t_t + t_p = \frac{D_i}{R_{u_i}^k} + \frac{DC_i}{c_i^j}
\] (3)

2.2. D2D communication between users

D2D is designed to reduce the load on the service base station by allowing user communication equipment to communicate directly within a range. Let the computing power of each free mobile terminal equipment be \( C_d \), \( c_i^d \) represents the computing power assigned to each user by the free equipment, indicated as \( c_i^d = \frac{c_i}{\sum_{u\in U} \mu_{u_i, r_u}} \). The maximum number of users per free equipment is \( Q^{ru_m} \), and the transmitting power per free mobile terminal is \( P_d \). For D2D pairs, sharing the same channel causes interference. In addition to D2D equipment sharing the channel, the channel also causes interference when users transmit data to the base station. User \( u_i \) offloads the task to the spare equipment over channel \( c_k \). We define the SNR received from the user \( u_i \) to the free equipment \( r_u \) using channel \( c_k \) as follows.

\[
\Gamma_{u_i, ru_m}^{c_k} = \frac{P_{u_i}^{c_k} \theta_{u_i, s_j}}{\sum_{u\in U, \forall u \neq u_i} \mu_{u_i, ru_m} r_{u_m} + \sum_{u\in U, s_j \neq s_j} \mu_{u_i, ru_m} r_{u_m} + C_d + \sum_{u\in U} \mu_{u_i, ru_m} + \sum_{u\in U} \mu_{u_i, ru_m} + \sigma^2}
\] (4)

The data transmission rate from user \( u_i \) using channel \( c_k \) to offload the task to \( ru_m \), can be expressed in Formula (5), and the service latency is shown in Formula (6).

\[
R_{u_i, ru_m}^{c_k} = B_c \log_2 (1 + \Gamma_{u_i, ru_m}^{c_k})
\] (5)

\[
T_{u_i, ru_m}^k = \frac{D_i}{R_{u_i, ru_m}^k} + \frac{DC_i}{c_i^m}
\] (6)

The benefits to the service provider

The service provider takes the reward \( O_i \) given by the service requester as revenue, which is proportional to the data size \( D_i \) of the service requester \( u_i \) and inversely proportional to the delay requirement \( T_i \), can be expressed in Formula (7), \( \chi \) represents a parameter whose unit is dollars per Mbps.

\[
O_i = \chi \frac{D_i}{T_i}
\] (7)
Each SBS and free equipment can provide computing services for multiple users according to their own computing capacity, and the revenue of SBS is \( P_{Sj} = \sum_{u_i \in U} \gamma^c_{u_i, s_j} O_l \), The benefit of free equipment can be expressed as \( P_{ru_m} = \sum_{u_i \in U} \mu^c_{u_i, ru_m} O_l \).

### 3. Problem formulation and solution

In the previous section, service delay and user cost were discussed. In order to better meet user requirements and maximize system cost performance, a new cost performance metric \( CP \) (Cost-Performance) was introduced, whose physical meaning was the price cost paid by each user to obtain the corresponding quality of service. It can be defined as the ratio between the average data rate per user and its price cost in Mbps/second/dollar. We define \( CP_{SYS} \) as the average cost performance of all users, can be expressed as follows.

\[
CP_{SYS} = \frac{\sum_{u_i \in U} CP_{u_i}}{N}
\]

(8)

Where, \( CP_{u_i} \) represents the cost performance of user \( u_i \), and is defined as follows.

\[
CP_{u_i} = \begin{cases}
\frac{D_i}{T^k_{u_i, s_j}} / O_l, & u_i \in U, \; u_i \notin P \\
\frac{D_i}{T^k_{u_i, ru_m}} / O_l, & u_i \in U, \; u_i \in P
\end{cases}
\]

(9)

Where \( P \) stands for the set of users that offload tasks to free equipment.

Thus, the computational resource allocation problem can be reduced to the following optimization model, as shown in Equation (10).

\[
P1: \text{maximize} \quad \sum_{u_i \in U} CP_{u_i} \\
P2: \text{s.t.} \quad \begin{cases}
C1: T^k_{u_i, s_j} \leq T^k_{u_i, ru_m} \leq T_i, \; u_i \in U, \; c_k \in K, \; s_j \in S \\
C2: \sum_{u_i \in U} \gamma^c_{u_i, s_j} \leq n^s_{\max}, \; \gamma^c_{u_i, s_j} \in \{0,1\} \\
C3: \sum_{u_i \in U} \mu^c_{u_i, ru_m} \leq E^d_{\max}, \; \mu^c_{u_i, ru_m} \in \{0,1\} \\
C4: \sum_{u_i \in U} \mu^c_{u_i, ru_m} P_{d \in c_i} \leq P_{d \max}, \; \mu^c_{u_i, ru_m} \in \{0,1\} \\
C5: \sum_{u_i \in U} \mu^c_{u_i, ru_m} E_{d \in c_i} \leq E_{d \max}, \; \mu^c_{u_i, ru_m} \in \{0,1\}
\end{cases}
\]

(10)

Where, C1 indicates that the delay of the user should meet the delay requirements that the user can receive; C2 represents the maximum number of users that each base station can accommodate; C3 represents the constraint of the maximum number of users that the free equipment can accommodate; C4 represents the maximum power constraint that an free equipment can provide; C5 represents the traffic constraints that free equipment can provide.

Obviously, the optimization problem in Equation (10) is a NP hard problem, and it is difficult to obtain the optimal solution. Therefore, this paper designs an improved Gale-Shapley algorithm based on the algorithm of matching game to obtain the feasible solution of this problem. First, the matching of service requester to service provider is modeled as a many-to-one matching. Service requesters and service providers can be regarded as two opposing participants of selfishness and rationality, both trying to improve their own interests in the matching process. A matching game algorithm is used to find a stable match between the service requester and the service provider. Stable matching is the optimal match for the initiator, a match acceptable to all matching participants. Stable matching means...
the robustness of preferences, which is good for both participants. In unstable matching, participants may exchange matching objects in order to improve their own interests.

Definition 1, define a matching M is a multivalued mapping, \( \forall u_i \in U, \forall R P^y_r \in RP \), we Define \( R P^y_r \) to represent resource r of resource type y, \( q_{RP^y_r} \) is the maximum number of users the resource \( R P^y_r \) can hold.

1. \( (u_i, R P^y_r) \in M \),
2. \( | M (u_i) | <= 1 \),
3. \( | M (RP^y_r) | <= q_{RP^y_r} \),
4. if and only if \( u_i \in M (RP^y_r) \), \( M (u_i) = RP^y_r \),

Definition 2, assume a matching M, which represents a match between a user and a resource \( (u_i, R P^y_r) \) \( \in M \). If there is no blocking pair (BP) in M, then a matching M is considered a stable match. A user-resource pair \( (u_i, R P^y_r) \) is called a BP if one of the following conditions is met.

1. The resource \( R P^y_r \) is under quota, and in user \( u_i \)'s priority list the resource \( R P^y_r \) is ahead of the resource that \( u_i \) is currently matching.
2. The resource \( R P^y_r \) quota is full, and the \( u_i \) position in the resource \( R P^y_r \) priority list is higher than any of the existing matching users, and in the \( u_i \) priority list, the \( R P^y_r \) ranking is higher than the \( u_i \) currently matching resource.

In order to find a stable match. First, establish a list of preferences for each other between the user and the service provider (base station, free equipment), denoted as \( L^{u_i, i} \) \( u_i \), \( L^{u_i, j} \) \( u_j \). In the process of establishing the preference list, delay constraints and signal-to-noise ratio requirements need to be considered. The set of resources that satisfy the user delay constraint and the signal-to-noise ratio constraint is called the acceptable set. After finding the acceptable set of all users, the resources are sorted in descending order according to their preference. For users, they tend to delay less resources. To facilitate calculations, users sharing the same resource share the spectrum bandwidth and CPU rate, and the CPU rate allocated to each user is only related to the number of users sharing the resource. However, before the matching is completed, the number of users sharing resources is unknown. We define the capacities of SBSs \( s_j \) and free equipment \( r_u m \) as \( Q_j \) and \( Q^m \). In order to calculate potential service delay, assume that computing resources allocated to each user to \( \frac{1}{Q} \) means resources can accommodate a maximum users.

In this article, consider potential service delays when calculating user and resource preferences, expressed as \( T^{u_i, s_j}_{i,j} \) and \( T_{u_i, u_j} \). The preference list is calculated as follows.

\[
L^{s_j}_{u_i} = T^{u_i}_{u_i, i} \cdot s_j = \frac{D_i}{R_{t_{ij}}} + \frac{D_c_i}{c_{i}^{i}} + \frac{D_i}{R_{t_{ij}}} + \frac{D_c_i}{c_{i}^{i}}
\]

\[
L^{r_u_m}_{u_i} = T^{u_i}_{u_i, r_u_m} = \frac{D_i}{R_{t_{ij}}^{r_u_m}} + \frac{D_c_i}{c_{i}^{r_u_m}} + \frac{D_i}{R_{t_{ij}}^{r_u_m}} + \frac{D_c_i}{c_{i}^{r_u_m}}
\]

Where \( R_{t_{ij}}^{c_k} \) represents the data transfer rate of user \( u_i \) using channel \( c_k \) to offload tasks to base station \( s_j \), is represented as \( R_{t_{ij}}^{c_k} = B_{c_k} \log_2 (1 + \frac{P_{u_i} d_{u_i, s_j}}{\sigma^2}) \), where \( R_{t_{ij}}^{c_k} \) represents the data transfer rate of user \( u_i \) using channel \( c_k \) to offload tasks to free equipment \( r_u_m \), is represented as \( R_{t_{ij}}^{c_k} = B_{c_k} \log_2 (1 + \frac{P_{u_i} d_{u_i, s_j}}{\sigma^2}) \).

When service providers express their preferences to users, they should also consider service delay. They pursue to provide short service time for each user so that more users can be served and get more remuneration. Therefore, the service provider's preference for users is based on the rate of price to service delay, can be expressed as \( L^{u_i}_{s_j} = \frac{O_i}{T^{u_i}_{u_i, s_j}} \) and \( L^{u_i}_{r_u_m} = \frac{O_i}{T^{u_i}_{u_i, r_u_m}} \).
Algorithm 1: Modified Gale-Shapley Algorithm.

Input: U, RP₁, RP₂ (base station, free equipment), Lᵢᵤ, Lᵢᵤᵐ, Lᵢₛ, Lᵢₛᵣᵣ;
Output: Matching M;
Initialization: set M empty, set all user free, \text{L}^\text{strun}_i = \text{L}^\text{strun}_i \cup \text{L}^\text{strun}_i;
1: while some user \text{u}_j is free and \text{u}_j has a non-empty preference list do
2: for all \text{u}_j \in U do
3: \text{u}_j proposes to the first entity \text{RP}^\text{strun}_i in \text{L}^\text{strun}_i, and then remove \text{RP}^\text{strun}_i from \text{L}^\text{strun}_i;
4: \text{M} = \text{M} \cup (\text{u}_j, \text{RP}^\text{strun}_i);
5: end for
6: for all \text{RP}^\text{strun}_i, \text{RP}^\text{strun}_i \in \text{RP}^1 do
7: While \text{RP}^\text{strun}_i is over-subscribed do
8: find the worst pair (\text{u}_{\text{w}}^\text{strun}, \text{RP}^\text{strun}_i) assigned to \text{RP}^\text{strun}_i in its candidate list;
9: \text{M} = \text{M} \setminus (\text{u}_{\text{w}}^\text{strun}, \text{RP}^\text{strun}_i);
10: end while
11: end for
12: for all \text{RP}^\text{run}_i, \text{RP}^\text{run}_i \in \text{RP}^2 do
13: While \text{RP}^\text{run}_i is over-subscribed do
14: find the worst pair (\text{u}_{\text{w}}^\text{run}, \text{RP}^\text{run}_i) assigned to \text{RP}^\text{run}_i in its candidate list;
15: \text{M} = \text{M} \setminus (\text{u}_{\text{w}}^\text{run}, \text{RP}^\text{run}_i);
16: end while
17: end for
18: end while
19: Terminate with a matching M.

After the preference list is set, algorithm 1 is applied to obtain the effective match between users and resources. Algorithm 1 is based on Gale-Shapley algorithm, so the total computational complexity is O(m), where m is the total length of the preference list.

4. Performance evaluation
Consider a three-layer heterogeneous cellular network consisting of one MBS, multiple SBS, and multiple mobile equipment. The MBS is deployed in the center of the network with a coverage radius of 500m. A number of SBS with MEC servers are randomly distributed in the network, and mobile equipment are also randomly distributed in the network. Mobile equipment include free equipment and mobile users. The specific simulation experiment parameters are shown in Table 1.

| Table 1. Simulation parameters |
|-------------------------------|----------------|
| parameters | value |
| U | [10,160] |
| S | 5 |
| RU | 20 |
| B | 5MHz |
| Q° | 15 |
| Q° | 3 |
| T | [6,7]s |
| D | [1.2]Mb |
| C | [5.6] * 10⁶ cycles/s |
| C | [2] * 10⁴ cycles/s |
| K | 10 |
We evaluate the proposed algorithm (Modified Gale-Shapley Algorithm) and the random resource allocation algorithm e.r.t user’s service latency, user’s satisfaction, total revenue from SBS and free equipment and the system utility function. Figure 2 shows the user’s average service delay of the improved Gale-Shapley algorithm and random method. As the number of users increases, the average service delay of users also increases. Since this paper is designed to share computing resources among users who offload to the same SBS or free equipment. As the number of user’s increases, the computing resources obtained by each user will decrease and the user service delay will also increase. However, it can be seen from the figure that the performance of the proposed algorithm is better than that of the random method. When the number of users is less than 40, the average service delay of the random method is lower. Because In the algorithm proposed in this article, all users choose their preferred resources. Therefore, the number of users sharing resources is too large, leading to the increase of service delay. When the number of users is greater than 40, the user performance of the algorithm proposed in this paper is significantly better than the random algorithm. That's because in the algorithm proposed in this paper, after users apply to resources, resources leave suitable users based on their preference list and capacity. However, in the random method, when a user sends an application to a resource, the resource can only passively receive the user.

![Figure 2. Average service latency](image1.png)

![Figure 3. User’s satisfaction](image2.png)

Figure 3 shows the user’s satisfaction. The user’s satisfaction of the two methods decreases with the increase of the number of service request users. When the number of service request users in the system is small, as the number of service request users increases, the user satisfaction rate of the random method decreases rapidly. The improved Gale-Shapley algorithm proposed in this paper has a relatively flat decline in user satisfaction. When the number of service request users reaches 160, the user satisfaction rate of the random method is reduced to 28%. The user satisfaction of the algorithm proposed in this paper can still reach 65%.

From the service provider's perspective, they derive benefits from the service request user. Figure 4 shows how the revenue earned by the service provider varies with the number of service requester users. In this paper, the user will pay only if the actual delay is less than the user's delay requirement.
The revenue of the service provider depends on the number of matched users. When the network capacity is not reached, almost all users can match a resource. In the random method, there may be multiple users matching one resource, and there will be situations where the user's delay needs cannot be met. The algorithm proposed in this paper basically guarantees that each user can meet the delay demand within the capacity range. When users exceed network capacity, users need to compete for resources. And users with strict delay requirements offer higher prices, and they are more likely to be left behind by small base stations or idle devices. Therefore, the total revenue of small cells and idle equipment will continue to rise.

We use the CP to assess the performance of the system, as shown in figure 5. When the number of users is less than 40, the random method is better than the algorithm proposed in this paper. That's because users first send requests to their preferred resources, and some good resources may receive more requests than others. Therefore, when the number of users is small and the resources are sufficient, good resources that fully match the users may not be as good as those that have enough free resources. When the number of users is greater than 40, the algorithm in this paper is better than the random algorithm. When the number of users is less than 150, as the number of users increases, the performance decreases. That's because as the number of users increases, the resources shared by each user will decrease, which reduces the average system performance.

5. Conclusion
In this article, the problems of resource allocation have been studied in heterogeneous cellular networks. First, we proposed a three-layer resource allocation model under 5G heterogeneous cellular networks. Secondly, the resource allocation problem between users and resources was formulated as a multi-user resource allocation optimization problem with the goal of maximizing system energy efficiency. Then a resource allocation algorithm based on matching game is designed. Finally, we used MATLAB for simulation experiments. The simulations results have demonstrated that our proposed framework can provide distributive, close-to-optimal performance from both the users’ perspective and the system’s view.

Acknowledgments
This work was supported by the National Natural Science Foundation of China (No. 61662007).

References
[1] Pham, Q. V., Le Anh, T., Tran, N. H., Park, B. J., Hong, C. S. (2018) Decentralized computation offloading and resource allocation for mobile-edge computing: a matching game approach. J. IEEE Access, 6: 75868-75885.
[2] Zhang, J., Xia, W., Yan, F., & Shen, L. (2018) Joint computation offloading and resource allocation optimization in heterogeneous networks with mobile edge computing. J. IEEE Access, 6:19324-19337.

[3] Tang, L., & Hu, H. (2020) Computation Offloading and Resource Allocation for the Internet of Things in Energy-Constrained MEC-Enabled HetNets. J. IEEE Access, 8: 47509-47521.

[4] Xu, C., Zheng, G., Tang, L. (2020) Energy-Aware User Association for NOMA-Based Mobile Edge Computing Using Matching-Coalition Game. J. IEEE Access, 8: 61943-61955.

[5] Zheng, G., Xu, C., Tang, L. (2020) Joint User Association and Resource Allocation for NOMA-Based MEC: A Matching-Coalition Approach. In: 2020 IEEE Wireless Communications and Networking Conference (WCNC). South Korea. pp. 1-6.

[6] Liu, G., Zhao, H., Li, D. (2017) Resource allocation in heterogeneous networks: A modified many-to-one swap matching. In: 2017 IEEE 17th International Conference on Communication Technology (ICCT). Chengdu. pp. 508-512.

[7] Zhang H., Xiao Y., Bu S., et al. (2017) Computing resource allocation in three-tier IoT fog networks: A joint optimization approach combining Stackelberg game and matching. J. IEEE Internet of Things Journal, 4(5): 1204-1215.

[8] Jia, B., Hu, H., Zeng, Y., Xu, T., & Yang, Y. (2018) Double-matching resource allocation strategy in fog computing networks based on cost efficiency. J. Journal of Communications and Networks, 3: 237-246.

[9] Sardellitti, S., Scutari, G., Barbarossa, S. (2015) Joint optimization of radio and computational resources for multicell mobile-edge computing. J. IEEE Transactions on Signal & Information Processing Over Networks, 1(2): 89-103.

[10] Chen, X., Jiao, L., Li, W., Fu, X. (2015) Efficient multi-user computation offloading for mobile-edge cloud computing. J. IEEE Transactions on Networking, 24 (5): 2795-2808.